Appendix

The Appendix is organized as follows:

- Appendix A provides detailed setup and hyperparameters for experiments.
- Appendix B provides additional experiment results.
- Appendix C provides the generalization analysis of PerAda and the full proofs for Theorem 1 and Theorem 2.
- Appendix D provides the convergence analysis of PerAda and the full proofs for Theorem 3 and Theorem 4.

A. Experimental Details

A.1. Datasets and Model

Table 5. Summary of datasets.

Dataset	Task	# Training Samples	# Test Samples	# Validation Samples	# Clients	Data Partition	# Classes
CIFAR-10	image classification	45000	10000	5000	20	label-shift non-IID (synthetic)	10
Office-Home	image classification	12541	1656	1391	4	covariate-shift non-IID (nature)	65
CheXpert	multi-label image classification	180973	20099	22342	20	label-shift non-IID (synthetic)	5

FL datasets We summarize our FL datasets in Tab. 5.

- **CIFAR-10** [28] contains nature images for 10 classes, such as cat, bird, dog. We simulate label non-IID on CIFAR-10 using Dirichlet distribution $\text{Dir}(\alpha)$ [23] with $\alpha = 0.1$, creating different local data size and label distributions for $M = 20$ clients.
- **Office-Home** [61] contains images from four domains, i.e., Art, Clipart, Product, and Real Word. All domains share the same 65 typical classes in office and home. We simulate the feature non-IID by distributing the data from 4 domains to 4 clients, respectively [58].
- **CheXpert** [24] is a dataset of chest X-rays that contains 224k chest radiographs of 65,240 patients, and each radiograph is labeled for the presence of 14 diseases as positive, negative, and uncertain. We map all uncertainty labels to positive (U-Ones [24]). We follow the original CheXpert paper to report the AUC score as a utility metric on five selected diseases, i.e., Cardiomegaly, Edema, Consolidation, Atelectasis, Pleural Effusion. To create the label-shift non-IID on CheXpert, we view each possible multi-class combination as a “meta-category” and group all combinations that have less than 2000 training samples into a new meta-category, which results in a total of 19 meta-categories. Then we use Dirichlet distribution $\text{Dir}(\alpha)$ with $\alpha = 0.3$ to create label-shift non-IID based on the 19 meta-categories for $M = 20$ clients. Such FL data partition simulates a scenario where different hospitals (clients) have different majority diseases among their patients. Note that such meta-categories are only used to create FL non-IID data partition, and our utility metric AUC score is always calculated based on the five diseases, i.e., a 5-label image classification task.

The number of samples for each dataset is shown in Tab. 5, where we use a ratio of 9:1 to split the original training data into training data and validation data for each dataset.

Distillation datasets We summarize our out-of-domain distillation dataset as below:

- CIFAR-10: we use 50k (unlabeled) samples from the CIFAR-100 training dataset.
- Office-Home and CheXpert: we use 50k (unlabeled) samples from the CIFAR-10 training dataset.

In Figure 4, we conduct the ablation study of distillation on CIFAR-10.

- Distillation steps: we fix the distillation data fraction as 1 and increase steps.
- Distillation data fraction: we fix the distillation steps as 100 and increase the data fraction.
- Distillation datasets: we fix the distillation steps as 100, data fraction as 1, and use different distillation datasets. Specifically, we use 100.5k samples from the STL-10 unlabeled+training dataset, 50k samples from the CIFAR-100 training dataset, and 5k samples from the CIFAR-10 validation dataset.

Evaluation datasets As mentioned in Sec. 7.1, we evaluate pFL accuracy mainly under two metrics: Local-test (i.e., clients' corresponding local test data) and Global-test (i.e., the union of clients' local test data), to study the *personalized performance* and *generalization* (against label or covariate shifts), respectively. In addition, for CIFAR-10, we evaluate pFL generalization against distribution shifts on CIFAR-10.1 [51] and CIFAR-10-C [19]. CIFAR-10.1 contains roughly 2,000 new test images that share the same categories as CIFAR-10, and the samples in CIFAR-10.1 are a subset of the TinyImages dataset [60].

CIFAR-10-C [19] is natural corruption benchmark for test-time distribution shits, containing common image corruptions such as Blur, Gaussian Noise, and Pixelate. It is generated by adding 15 common corruptions plus 4 extra corruptions to the test images in the CIFAR-10 dataset.

Model We use a ResNet-18 [18] pretrained on ImageNet-1K [53] for all tasks. We additionally evaluate Office-Home on ResNet-34 [18] pretrained on ImageNet-1K. The pretrained models are downloaded from PyTorch [47].

Tab. 6 and Tab. 7 show the detailed model architectures of ResNet-18 and ResNet-34 model used for personalization on Office-Home, respectively. We use the number of parameters in the corresponding layers and the number of parameters in the full model to calculate the total number of # trainable parameters for different full model pFL and partial model pFL in Figure 1.

Since we use ResNet-18 for all datasets, the number of parameters of different kinds of layers for CIFAR-10 and CheXpert are the same, except for the output layer. This is because different datasets have different numbers of classes, which decide the size of the output layer. In Tab. 1, we report the parameters of the ResNet-18 model on CIFAR-10, where the output layer consists of 0.0051M parameters.

Table 6. Summary of model architectures of ResNet-18 model used for personalization on Office-Home.

Type	Detailed layers	# Params. in the layers
Full model	full model	11.21 M
Input layer	1st Conv. layer	4.73 M
Feature extractor	the model except last fully connected layer	11.16 M
Batch norm	batch normalization layers	0.0078M
Output layer	last fully connected layer	0.033 M
Adapter	residual adapter modules	1.44 M

Table 7. Summary of model architectures of ResNet-34 model used for personalization on Office-Home.

Type	Detailed layers	# Params. in the layers
Full model	full model	21.32 M
Input layer	1st Conv. layer	9.78 M
Feature extractor	the model except last fully connected layer	11.16 M
Batch norm	batch normalization layers	0.015 M
Output layer	last fully connected layer	0.033 M
Adapter	residual adapter modules	2.57 M

A.2. Training Details

We tuned the hyperparameters according to the personalized performance evaluated on the local validation data. We use SGD as the client optimizer. For each baseline method as well as our method, we tuned the (client) learning rate via grid search on the values {5e-4,1e-3, 5e-3, 1e-2} for CIFAR-10 and CheXpert, and {5e-4, 1e-3, 5e-3, 1e-2, 5e-2} for Office-Home. For PerAda, we use Adam as the server optimizer. We tuned the server learning rate via grid search on the values {1e-5,1e-4, 1e-3, 1e-2} for all datasets. The strength of regularization λ is selected from {0.1, 1} following [33] and we use the same λ for PerAda, Ditto, pFedMe. For pFedMe, we use the inner step of $K = 3$ as suggested in [59]. For APFL, the mixing parameter α is selected from {0.1, 0.3, 0.5, 0.7}. The final hyperparameters we used for PerAda are given in Tab. 8.

A.3. Experimental Setups for DP Experiments

Since the batch normalization layer in ResNet-18 requires computing the mean and variance of inputs in each mini-batch, creating dependencies between training samples and violating the DP guarantees, it is not supported in differentially private models. Thus, we turn to conduct DP experiments with a ViT-S/16-224 model [64], which is pretrained on ImageNet-21k [53]. We download the pretrained model from Hugging Face [63].

Following [42], we consider full client participation and perform local training with DP-SGD [1] for personalized models and the global model. On CIFAR-10, the local epoch is 1, and we run all methods for 10 communication rounds. We tuned the (client) learning rate via grid search on the values {0.01, 0.05,0.1, 0.2, 0.3 } for Ditto, PerAda w/o KD, and PerAda. The optimal learning rate for Ditto, PerAda w/o KD, and PerAda are 0.05, 0.1, and 0.2, respectively. For PerAda, we set

Table 8. Hyperparameters of PERADA for each dataset.

Hyperparameter	CIFAR-10	Office-Home	CheXpert
Batch size	64	128	256
Clients per round	8	4	8
Local epochs	10	1	1
# training rounds	200	100	30
Regularization strength λ	1	0.1	1
Client learning rate	0.01	0.05	0.01
Server learning rate	1e-3	1e-4	1e-5
Distillation step	500	100	50
Distillation batch size	2048	256	128

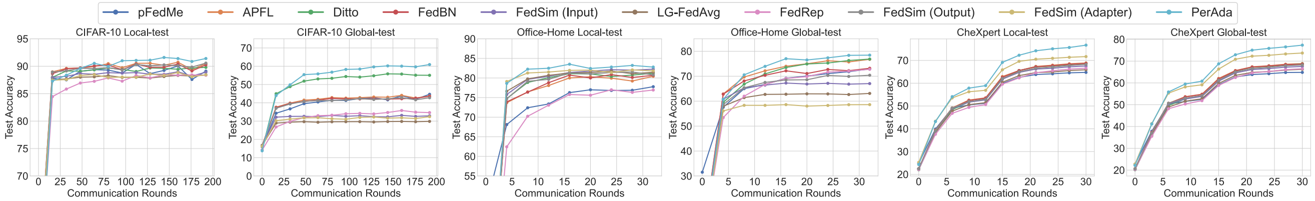

Figure 6. Averaged test accuracy of personalized models from participating clients at each communication round.

the distillation batch size as 32. We select the sever learning rate from {0.005, 0.003, 0.001}, and the optimal server learning rate is 0.005.

We set the DP parameter $\delta = 10^{-5}$ and evaluate the averaged pFL accuracy under $\mathrm{Local\text{-}test}$. We set the noise level σ as $0.8, 1, 1.5$ for DP-SGD training to obtain the privacy budgets $\epsilon = 5.99 \pm 3.03, 3.7 \pm 2.12, 1.81 \pm 1.12$ used in Tab. 4, respectively. Under each privacy budget, we tuned the clipping threshold via grid search from {1, 2, . . . , 10 } for each method.

B. Additional Experimental Results and Analysis

In this section, we provide additional experimental results and analysis, including (1) Convergence analysis; (2) analysis of pFL performance under different model architectures Office-Home; (3) pFL performance under different data heterogeneity degrees on CheXpert; (4) generalization comparison of the global model of different pFL methods; (5) effect of the pretrained model; (6) effect of regularization strength λ.

Convergence We present the learning performance from the convergence perspective in Figure 6, where we report the averaged test accuracy of personalized models from the participated clients at each communication round. It shows that PERADA achieves the best convergence speed and converges to a higher personalized performance (local-test) and generalization performance (global-test).

Performance under different model architectures (ResNet-18 and ResNet-34) on Office-Home. Figure 1 shows the performance of different pFL under ResNet-18 and ResNet-34. Cross different network architecture, PERADA is able to achieve the best personalized performance and generalization with the fewest number of trainable parameters. For larger model, the number of updated parameters difference between full model personalization and our adapter personalization will be larger, reflecting our efficiency.

Performance under different data heterogeneity degrees on CheXpert. Tab. 3 shows under different data heterogeneity degrees $\mathrm{Dir}(1)$ and $\mathrm{Dir}(0.3)$ on CheXpert, PERADA achieves the best personalized performance and generalization. It also verifies that adapter-based personalization methods, including FEDALT, FEDSIM, PERADA are especially effective on the X-ray data CheXpert.

Generalization comparison of the global model of different pFL methods. Tab. 2 compare the generalization performance of the global model in our method to the global model in other full model pFL methods (PFEDME, APFL, DITTO) and generic FL methods (FEDAVG, FEDPROX [32], FEDDYN [2], FEDDF [39]) on CIFAR-10. MTL and partial model pFL methods are

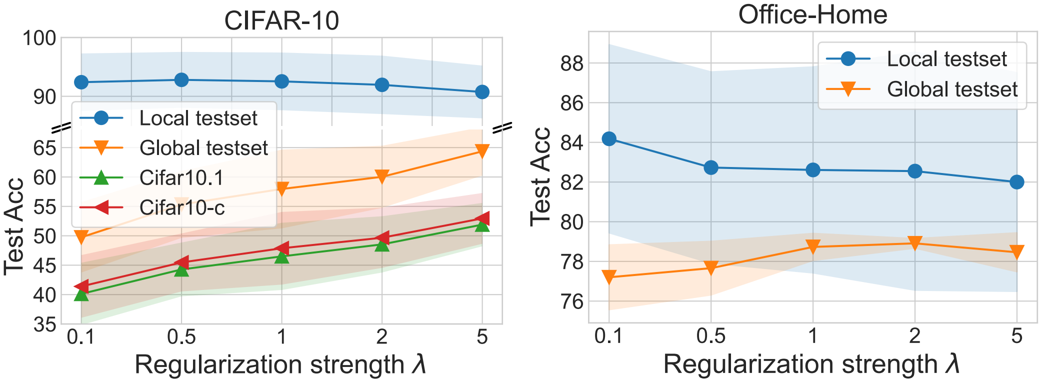

Figure 7. Effect of λ on PERADA on CIFAR-10 and Office-Home.

excluded from the compression because they do not train a complete global model. We use the same distillation dataset and distillation steps and data size for FEDDF and PERADA to ensure a fair comparison.

The results show that the global model of PERADA outperforms these baselines, which verifies that KD improves our global model, and the improved performance of personalized models is due to a well-generalized global model.

Effect of pretrained models. Starting personalization from a pretrained model, such as FEDAVG model [44, 48], is common in pFL, so we report the results with FEDAVG pretrained model (on FL data from scratch) for all methods[3] on CIFAR-10. The results in Figure 5 show that PERADA also achieves comparable personalized performance and higher generalization than baselines with FEDAVG pretrained model. Moreover, Theorem 1 shows that high-quality local models (enabled by good pretrained model) can further improve generalization. Here, we use ImageNet as an example of high-quality pretrained models, which leads to even higher personalized performance and generalization for PERADA. Additionally, pretrained models lead to significantly higher pFL accuracy than random initialization for all existing methods; therefore, leveraging a pretrained model, which is often available for modern deep neural networks [5], is practical and beneficial not only for PERADA but also for existing pFL methods.

Effect of λ. Results on CIFAR-10 and Office-Home in Figure 7 shows that moderately increasing regularization strength λ can improve generalization, but it also degrades the personalized performance, which matches the observation for ℓ_2 regularization-based pFL methods in [48].

[3] FEDSIM is omitted here because its results are similar to FEDALT [48]

C. Generalization Analysis

We give the discussions and analysis for our generalization bounds. The outline of this section is as follows:

- Appendix C.1 provides more discussions on Theorem 1.
- Appendix C.2 provides the peliminaries for generalization bounds and introduces several useful lemmas.
- Appendix C.3 provides the proofs for generalization bounds of global model in Theorem 1.
- Appendix C.4 provides the proofs for generalization bounds of personalized model in Theorem 2.

C.1. Additional Discussion

Additional Discussion on Theorem 1. From Theorem 1, we can have the additional observations: (i) **Client heterogeneity.** Larger heterogeneity, i.e., higher distribution divergence $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathbb{D}_m, \mathbb{D})$ between local and global datasets, could undermine the generalization of g, echoing the implications in [39, 68] (ii) **Number of classes**. The smaller number of classes k is favorable to generalization, as the classification task with fewer classes is easier to learn. We note that previous FL generalization bounds [39, 44, 68] are limited to binary classification cases.

C.2. Peliminaries for Generalization Bounds

Here we introduce several existing definitions and lemmas from learning theory.

Lemma 1 (Empirical Rademacher complexity [55]). *$\mathcal{G}$ be a set of functions $\mathcal{Z} \to [a, b]$, $\forall\delta > 0$. Let $Z_1, \ldots, Z_n$ be i.i.d. random variables on $\mathcal{Z}$ following some distribution P. The empirical Rademacher complexity of $\mathcal{G}$ with respect to the sample $(Z_1, \ldots, Z_n)$ is*

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) := \mathbb{E}_\sigma \left[\sup_{g\in\mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(x_i) \right] \tag{1}$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)^\top$ with $\sigma_i \sim \operatorname{unif}\{-1, 1\}$, which is are known as Rademacher random variables.

Moreover, with probability at least $1 - \delta$, we have w.r.t the draw of S that

$$\forall g \in \mathcal{G}, \mathbb{E}[g(\mathcal{Z})] \leq \frac{1}{n} \sum_{i=1}^{n} g(x_i) + 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3(b-a)\sqrt{\frac{\log(2/\delta)}{2n}} \tag{2}$$

Definition 1 (Risk [4]). We define a domain as a pair consisting of a distribution μ_S on inputs $\mathcal{X}$ and a labeling function $h^*_S : \mathcal{X} \to \Delta^k$. The probability according to the distribution μ_S that a hypothesis h disagrees with a labeling function h^*_S (which can also be a hypothesis) is defined as

$$\varepsilon_{\mu_S}(h) = \varepsilon_{\mu_S}(h, h^*_S) = \mathbb{E}_{(x,y)\sim\mu_S} |h(x)_y - h^*_S(x)_y| \tag{3}$$

Definition 2 ($\mathcal{H}$ -divergence [4]). Given a domain $\mathcal{X}$ with μ and μ' probability distributions over $\mathcal{X}$, let $\mathcal{H}$ be a hypothesis class on $\mathcal{X}$ and denote by $I(h)$ the set for which $h \in \mathcal{H}$ is the characteristic function; that is, where $(x, y) \in I(h) \Leftrightarrow h(x)_y = 1$. The $\mathcal{H}$ -divergence between μ and μ' is

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mu, \mu') = 2 \sup_{h\in\mathcal{H}} |\Pr_{\mu}(I(h)) - \Pr_{\mu'}(I(h))| \tag{4}$$

Lemma 2 (Domain adaptation [4]). *Let $\mathcal{H}$ be a hypothesis space on $\mathcal{X}$ with VC dimension d. Considering the distributions μ_S and μ_T. If $\mathcal{D}'_S$ and $\mathcal{D}'_T$ are samples of size n from μ_S and μ_T respectively and $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}'_S, \mathcal{D}'_T, n)$ is the empirical $\mathcal{H}$ -divergence between samples, then for every $h \in \mathcal{H}$ and any $\delta \in (0, 1)$, with probability at least $1 - \delta$ (over the choice of samples) , there exists,*

$$\varepsilon_{\mu_T}(h) \leq \varepsilon_{\mu_S}(h) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}'_S, \mathcal{D}'_T) + 4\sqrt{\frac{2d\log(2n) + \log(2/\delta)}{n}} + \lambda$$

where $\lambda = \varepsilon_{\mu_T}(h^*) + \varepsilon_{\mu_S}(h^*)$ and $h^* := \arg\min_{h\in\mathcal{H}} \varepsilon_{\mu_T}(h) + \varepsilon_{\mu_S}(h)$ corresponds to ideal joint hypothesis that minimizes the combined error.

C.2.1 Useful Lemmas

Then, we introduce several useful lemmas.

Lemma 3. *[22] For any $v \in \mathbb{R}^k$ and $y \in [k]$,*

$$2\left(1-v\right)_y \geq \mathbb{1}\left[y \neq \arg\max_i v_i\right].$$

Proof. Let $v \in \mathbb{R}^k$ be given, and consider two cases. For the first case, if $y = \arg\max_i v_i$, then $v \in [0,1]^k$ implies $2(1-v) \geq 0 = \mathbb{1}\left[y \neq \arg\max_i v_i\right]$. For the second case, if $y \neq \arg\max_i v_i$, then $v_y \leq 1/2$ and $2(1-v) \geq 1 = \mathbb{1}\left[y \neq \arg\max_i v_i\right]$. Combining the two cases together, we prove the lemma. □

Lemma 4. *For any functions $\mathcal{H}$ with $\mathcal{H} \ni h : \mathcal{X} \to \mathbb{R}^k$, since $\mathcal{H}$ takes values in $\mathbb{R}^k$, let $H|_j$ denote the Rademacher complexity of each class j,*

$$\mathrm{Rad}_n\left(\{(x,y) \mapsto 1 - h(x)_y : h \in \mathcal{H}\}\right) = \mathcal{O}\left(\sqrt{k}\max_j \mathrm{Rad}_n(\mathcal{H}|_j)\right)$$

where $\max_k \mathrm{Rad}_n(\mathcal{H}|_k)$ is the worst-case per class Rademacher complexity.

Proof. The proof follows from a multivariate Lipschitz composition lemma for Rademacher complexity due to [13, Theorem 1]; it remains to prove that $v \mapsto \psi(v)_y$ is 1-Lipschitz with respect to the ℓ_∞ norm for any $v \in \mathbb{R}^k$ and $y \in [k]$.

$$\frac{\mathrm{d}}{\mathrm{d}v_y}\psi(v)_y = \frac{\mathrm{d}}{\mathrm{d}v_y}(1-v_y) = -1, \quad \frac{\mathrm{d}}{\mathrm{d}v_{i\neq y}}\psi(v)_y = \frac{\mathrm{d}}{\mathrm{d}v_{i\neq y}}(1-v_y) = 0$$

and therefore $\|\nabla\psi(v)_y\|_1 = 1$ and thus, by the mean value theorem, for any $u \in \mathbb{R}^k$ and $v \in \mathbb{R}^k$, there exists $z \in [u,v]$ such that

$$|\psi(v)_y - \psi(u)_y| = |\left\langle \nabla\psi(z)_y, v-u\right\rangle \leq \|v-u\|_\infty \|\nabla\psi(v)_y\|_1 \leq \|v-u\|_\infty.$$

In particular, $v \mapsto \psi(v)_y$ is 1-Lipschitz with respect to the ℓ_∞ norm. Applying the aforementioned Lipschitz composition rule [13, Theorem 1],

$$\mathrm{Rad}_n\left(\{(x,y) \mapsto 1 - h(x)_y : h \in \mathcal{H}\}\right) = \mathrm{Rad}_n\left(\{(x,y) \mapsto \psi(h(x))_y : h \in \mathcal{H}\}\right) = \mathcal{O}\left(\sqrt{k}\max_j \mathrm{Rad}_n(\mathcal{H}|_j)\right)$$

□

Lemma 5. *For any functions $\mathcal{H}_m$ with $\mathcal{H}_m \ni h_m : \mathcal{X} \to \mathbb{R}^k$ with any $m \in [M]$, and $h \in \mathcal{H}$ where $h(x) := \frac{1}{M}\sum_{m=1}^M h_m(x)$ for any $x \in \mathcal{X}$*

$$\mathrm{Rad}_n(\mathcal{H}|_j) = \frac{1}{M}\sum_{m=1}^M \mathrm{Rad}_n(\mathcal{H}_m|_j) \tag{5}$$

Proof.

$$\begin{aligned}
\mathrm{Rad}_n(\mathcal{H}|_j) &= \frac{1}{n}\mathbb{E}_\epsilon \sup_{h\in\mathcal{H}}\sum_{i=1}^n \epsilon_i h\left(x_i\right)_j \\
&= \frac{1}{n}\mathbb{E}_\epsilon \sup_{h_1,\ldots,h_M\in\mathcal{H}}\sum_{i=1}^n \epsilon_i\left(\frac{1}{M}\sum_{m=1}^M h_m(x_i)\right)_j \\
&= \frac{1}{M}\sum_{m=1}^M \frac{1}{n}\mathbb{E}_\epsilon \sup_{h_m\in\mathcal{H}}\sum_{i=1}^n \epsilon_i h_m(x_i)_j \\
&= \frac{1}{M}\sum_{m=1}^M \mathrm{Rad}_n(\mathcal{H}_m|_j)
\end{aligned}$$

□

C.3. Proofs for Generalization Bounds of Global Model Theorem 1

Overview Recall the definition of distillation distance:

$$\Phi_{\mu,n}(h_1,\dots,h_M;g) := \frac{1}{n}\sum_{i=1}^{n}\|g(x_i) - \frac{1}{M}\sum_{m=1}^{M} h_m(x_i)\|_1 \tag{6}$$

which measures the output difference between the global model and the ensemble of local models. The server distillation (Line 21 in Algorithm 1) essentially finds the global model g with a small distillation distance $\Phi_{\mu_{\text{aux}},n_{\text{aux}}}$, meaning that its outputs are close to the ensemble outputs of local models $f_1,\dots,f_M$ on the out-of-domain distillation dataset $\mathbb{D}_{\text{aux}}$.

For the generalization bounds of the global model, we aim to show g can have good generalization bounds on μ with KD if it (1) distills knowledge accurately from teachers $\{f_m\}$ and (2) the teachers $\{f_m\}$ performs well on their local distributions $\{\mu_m\}$. To sketch the idea, by Lemma 3, we can upper bound error probabilities of g with the expected distillation distances and errors of local models (i.e., teachers) on μ:

$$\Pr_{(x,y)\sim\mu}\left[\arg\max_{y'} g(x)_{y'} \neq y\right] = \mathbb{E}_{(x,y)\sim\mu}\mathbb{1}\left[\arg\max_{y'} g(x)_{y'} \neq y\right] \tag{7}$$

$$\leq 2\underbrace{\mathop{\mathbb{E}}_{x\sim\mu}\left\|g(x) - \frac{1}{M}\sum_{m=1}^{M} h_m(x)\right\|_1}_{\text{ensemble distillation distance}} + 2\underbrace{\mathop{\mathbb{E}}_{(x,y)\sim\mu}\left(1 - \frac{1}{M}\sum_{m=1}^{M} h_m(x)_y\right)}_{\text{errors of teacher models}} \tag{8}$$

Then, we can relate the errors of local models h_m on μ to μ_m with prior arts from domain adaptation [4].

To simply our notations, we define **"virtual hypothesis"** $h \in \mathcal{H} : \mathcal{X} \to [0,1]^k$ **, whose outputs are the ensemble outputs from all local models**:

$$h(x) := \frac{1}{M}\sum_{m=1}^{M} h_m(x).$$

Main Analysis Let us recall Theorem 1.

Theorem 1 (Generalization bound of PerAda global model)**.** *Consider empirical datasets* $\mathbb{D} \sim \mu, \mathbb{D}_{\text{aux}} \sim \mu_{\text{aux}}, \mathbb{D}_m \sim \mu_m$ *with* $|\mathbb{D}| = |\mathbb{D}_m| = n, |\mathbb{D}_{\text{aux}}| = n_{\text{aux}}$*. Let* d_m *be the VC dimension of* $\mathcal{H}_m$*,* $\text{Rad}_{n_{\text{aux}}}$ *be the empirical Rademacher complexity measured on* n_{aux} *samples. With probability at least* $1-\delta$*, for every* $h_m \in \mathcal{H}_m, \forall m \in [M]$ *and* $g \in \mathcal{G}$*, we have* $\Pr_{(x,y)\sim\mu}\left[\arg\max_{y'} g(x)_{y'} \neq y\right] \leq \mathop{2\mathbb{E}}_{(x,y)\sim\mu}[1 - g(x)_y] \leq \mathcal{O}(k^{3/2}[\max_j(\frac{1}{M}\sum_{m=1}^{M}\text{Rad}_{n_{\text{aux}}}(\mathcal{H}_m|_j)) + \max_j \text{Rad}_{n_{\text{aux}}}(\mathcal{G}|_j)]) + \frac{6}{M}\sum_{m=1}^{M}(\frac{4}{3}\sqrt{\frac{2d_m\log(2n)+\log(6M/\delta)}{n}} + \sqrt{\frac{\log(6M/\delta)}{2n}} + \sqrt{\frac{\log(6/\delta)}{2n_{\text{aux}}}} + \mathcal{O}(\text{Rad}_n(\mathcal{H}_m))) + \frac{1}{M}\sum_{m=1}^{M}(\underbrace{2\mathtt{ERR}(\mathbb{D}_m,h_m)}_{\textit{local empirical risk}} + \underbrace{\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathbb{D}_m,\mathbb{D})}_{\textit{client heterogeneity}} + \lambda_m) + 2\underbrace{\Phi_{\mu_{\text{aux}},n_{\text{aux}}}(h_1,\dots,h_M;g)}_{\textit{ensemble distillation distance}} + 4\underbrace{\mathbb{TV}(\mu,\mu_{\text{aux}})}_{\textit{TV divergence}}$*, where* $\mathtt{ERR}(\mathbb{D}_m,h_m) = \frac{1}{n}\sum_{j=1}^{n}\left[1 - h_m(x_{m,j})_{y_{m,j}}\right], \lambda_m = \varepsilon_{\mu_m}(h^*) + \varepsilon_\mu(h^*), h^* := \arg\min_{h\in\mathcal{H}} \varepsilon_{\mu_m}(h) + \varepsilon_\mu(h).$

To prove the generalization bounds of the global model Theorem 1, we use Lemma 6 as a bridge.

Lemma 6. *Let classes of bounded functions* $\mathcal{H}$ *and* $\mathcal{G}$ *be given with* $h \in \mathcal{H} : \mathcal{X} \to [0,1]^k$ *and* $g \in \mathcal{G} : \mathcal{X} \to [0,1]^k$*. Suppose* $\{x_i\}_{i=1}^{n_{\text{aux}}}$ *is sampled from a distribution* μ_{aux}*. For every* $h \in \mathcal{H}$ *and every* $g \in \mathcal{G}$*, with probability at least* $1-\delta$*,*

$$\mathbb{E}_{(x,y)\sim\mu} g(x)_y \leq \mathbb{E}_{(x,y)\sim\mu} h(x)_y + \frac{1}{n_{\text{aux}}}\sum_{i=1}^{n_{\text{aux}}}\min\{1, \|g(x_i) - h(x_i)\|_1\} + 2\mathbb{TV}(\mu,\mu_{\text{aux}}) + 3\sqrt{\frac{\log(2/\delta)}{2n_{\text{aux}}}}$$

$$+ 2\sum_{y'=1}^{k}\left(\text{Rad}_{n_{\text{aux}}}\left(\{x \mapsto h(x)_{y'} : h \in \mathcal{H}\}\right) + \text{Rad}_{n_{\text{aux}}}\left(\{x \mapsto g(x)_{y'} : g \in \mathcal{G}\}\right)\right)$$

Proof. To start, for any $h \in \mathcal{H}, g \in \mathcal{G}$, write

$$\mathbb{E}_{x,y} g(x)_y = \mathbb{E}_{x,y}(g(x) - h(x))_y + \mathbb{E}_{x,y} h(x)_y$$

For the first term, since $h : \mathcal{X} \to [0,1]^k$ and $g : \mathcal{X} \to [0,1]^k$, by Holder's inequality

$$\begin{aligned}
\mathbb{E}_{x,y}(g(x) - h(x))_y &= \int \min\left\{1, (g(x) - h(x))_y\right\} \mathrm{d}\mu(x,y) \\
&\leq \int \min\left\{1, \|g(x) - h(x)\|_1\right\} \mathrm{d}\mu_{\mathcal{X}}(x)
\end{aligned}$$

Here we need 1 in $\min\left\{1, (g(x) - h(x))_y\right\}$ to make the upper bound tighter, since $(g(x) - h(x))_y \leq 1$ always hold.

Note that for any two measures μ and ν, and for any continuous function $f(x)$ in $[0,1]$,

$$\begin{aligned}
\int h(x)(\mathrm{d}\mu(x) - \mathrm{d}\nu(x)) &= \int_{x \in A} f(x)(\mathrm{d}\mu(x) - \mathrm{d}\nu(x)) + \int_{x \in B} f(x)(\mathrm{d}\mu(x) - \mathrm{d}\nu(x)) \\
&\leq |\mu(A) - \nu(A)| + |\mu(B) - \nu(B)| \\
&\leq 2 \sup_{\text{measurable } S} |\mu(S) - \nu(S)| = 2\mathbb{TV}(\mu, \nu),
\end{aligned}$$

where $A = \{x : \mathrm{d}\mu(x) \geq \mathrm{d}\nu(x)\}$ and $B = \{x : \mathrm{d}\mu(x) < \mathrm{d}\nu(x)\}$.

Once again invoking standard Rademacher complexity arguments Lemma 1, with probability at least $1 - \delta$, every mapping $x \mapsto \min\left\{1, \|g(x) - h(x)\|_1\right\}$ where $h \in \mathcal{H}$ and $g \in \mathcal{G}$ satisfies

$$\begin{aligned}
\int \min\left\{1, \|g(x) - h(x)\|_1\right\} \mathrm{d}\mu_{\mathcal{X}}(x) &\leq \int \min\{1, \|g(x) - h(x)\|_1\} \mathrm{d}\mu_{\mathsf{aux}}(x) \\
&\quad + \int \min\{1, \|g(x) - h(x)\|_1\}(\mathrm{d}\mu(x) - \mathrm{d}\mu_{\mathsf{aux}}(x)) \\
&\leq \int \min\{1, \|g(x) - h(x)\|_1\} \mathrm{d}\mu_{\mathsf{aux}}(x) + 2\mathbb{TV}(\mu, \mu_{aux}) \\
&\leq \frac{1}{n_{\mathsf{aux}}} \sum_{i=1}^{n_{\mathsf{aux}}} \min\left\{1, \|g(x_i) - h(x_i)\|_1\right\} + 2\mathbb{TV}(\mu, \mu_{\mathsf{aux}}) + 3\sqrt{\frac{\log(2/\delta)}{2n_{\mathsf{aux}}}} \\
&\quad + 2\,\mathrm{Rad}_{n_{\mathsf{aux}}}\left(\left\{x \mapsto \min\left\{1, \|g(x) - h(x)\|_1\right\} : h \in \mathcal{H}, g \in \mathcal{G}\right\}\right)
\end{aligned}$$

For the final Rademacher complexity estimate, first note $r \mapsto \min\{1, r\}$ is 1-Lipschitz and can be peeled off, and we use the definition of the empirical Rademacher complexity (Lemma 1), thus

$$\begin{aligned}
&\mathrm{Rad}_{n_{\mathsf{aux}}}\left(\left\{x \mapsto \min\left\{1, \|g(x) - h(x)\|_1\right\} : h \in \mathcal{H}, g \in \mathcal{G}\right\}\right) \\
&\leq \mathrm{Rad}_{n_{\mathsf{aux}}}\left(\left\{x \mapsto \|g(x) - h(x)\|_1 : h \in \mathcal{H}, g \in \mathcal{G}\right\}\right) \\
&= \mathbb{E}_\epsilon \sup_{\substack{h \in \mathcal{H} \\ g \in \mathcal{G}}} \frac{1}{n_{\mathsf{aux}}} \sum_{i=1}^{n_{\mathsf{aux}}} \epsilon_i \left\|g\left(x_i\right) - h\left(x_i\right)\right\|_1 \\
&\leq \sum_{y'=1}^{k} \mathbb{E}_\epsilon \sup_{\substack{h \in \mathcal{H} \\ g \in \mathcal{G}}} \frac{1}{n_{\mathsf{aux}}} \sum_{i=1}^{n_{\mathsf{aux}}} \epsilon_i \left|g\left(x_i\right) - h\left(x_i\right)\right|_{y'} \\
&= \sum_{y'=1}^{k} \mathrm{Rad}_{n_{\mathsf{aux}}}\left(\left\{x \mapsto |g(x) - h(x)|_{y'} : h \in \mathcal{H}, g \in \mathcal{G}\right\}\right).
\end{aligned}$$

Since h and g have range $[0,1]^k$, then $(h - g)_{y'}$ has range $[-1,1]$ for every y', and since $r \mapsto |r|$ is 1-Lipschitz over $[-1,1]$, combining this with the Lipschitz composition rule for Rademacher complexity and also the fact that a Rademacher random vector $\epsilon \in \{\pm 1\}^n$ is distributionally equivalent to its coordinate-wise negation $-\epsilon$, then, for every $y' \in [k]$

$$
\begin{aligned}
&\mathrm{Rad}_{n_{\text{aux}}}\left(\{x \mapsto |g(x) - h(x)|_{y'} : h \in \mathcal{H}, g \in \mathcal{G}\}\right)\\
&\leq \mathrm{Rad}_{n_{\text{aux}}}\left(\{x \mapsto (g(x) - h(x))_{y'} : h \in \mathcal{H}, g \in \mathcal{G}\}\right)\\
&= \frac{1}{n_{\text{aux}}}\mathbb{E}_{\epsilon} \sup_{h\in\mathcal{H}} \sup_{g\in\mathcal{G}} \sum_{i=1}^{n_{\text{aux}}} \epsilon_i \left(h\left(x_i\right) - g\left(x_i\right)\right)_{y'}\\
&= \frac{1}{n_{\text{aux}}}\mathbb{E}_{\epsilon} \sup_{h\in\mathcal{H}} \sum_{i=1}^{n_{\text{aux}}} \epsilon_i h\left(x_i\right)_{y'} + \frac{1}{n_{\text{aux}}}\mathbb{E}_{\epsilon} \sup_{g\in\mathcal{G}} \sum_{i=1}^{n_{\text{aux}}} -\epsilon_i g\left(x_i\right)_{y'}\\
&= \mathrm{Rad}_{n_{\text{aux}}}\left(\{x \mapsto h(x)_{y'} : h \in \mathcal{H}\}\right) + \mathrm{Rad}_{n_{\text{aux}}}\left(\{x \mapsto g(x)_{y'} : g \in \mathcal{G}\}\right)
\end{aligned}
$$

□

Inspired by [22], we introduce Lemma 3 to tackle the error probability $\mathrm{Pr}_{(x,y)\sim\mu}\left[\arg\max_{y'} g(x)_{y'} \neq y\right]$.

Let us define $\psi(v) = 1 - v$. According to Lemma 3, we can derive the upper bound for $\mathrm{Pr}_{x,y}\left[\arg\max_{y'} g(x)_{y'} \neq y\right]$ in Theorem 1 as below

$$
\begin{aligned}
\mathbb{E}_{x,y}\psi(g(x))_y &= \mathbb{E}_{x,y}\left(1 - g(x)_y\right)\\
&\geq \frac{1}{2}\mathbb{E}_{x,y}\left[\mathbb{1}\left[\arg\max_{y'} g(x)_{y'} \neq y\right]\right]\\
&= \frac{1}{2}\mathrm{Pr}_{x,y}\left[\arg\max_{y'} g(x)_{y'} \neq y\right] \qquad (9)
\end{aligned}
$$

Then we will study the upper bound for $\mathbb{E}_{x,y}\psi(g(x))_y$ in Lemma 7.

Lemma 7. *Let classes of bounded functions $\mathcal{H}$ and $\mathcal{G}$ be given with $h \in \mathcal{H} : \mathcal{X} \to [0,1]^k$ and $g \in \mathcal{G} : \mathcal{X} \to [0,1]^k$. Let classes of bounded functions $\mathcal{H}_m$ be given with $h_m \in \mathcal{H}_m : \mathcal{X} \to [0,1]^k$, $\forall m \in [M]$. For every $h_m \in \mathcal{H}_m, \forall m \in [M]$, and for every $g \in \mathcal{G}$, with probability at least $1-\delta$,*

$$
\begin{aligned}
\mathbb{E}_{(x,y)\sim\mu}[1 - g(x)_y)] \leq & \mathbb{E}_{(x,y)\sim\mu}[1 - h(x)_y] + \Phi_{\mu_{\text{aux}},n_{\text{aux}}}(h_1,\ldots,h_M; g) + 2\mathbb{T}V(\mu,\mu_{\text{aux}}) + 3\sqrt{\frac{\log(2/\delta)}{2n_{\text{aux}}}}\\
&+ \mathcal{O}\left(k^{3/2}\left[\max_j\left(\frac{1}{M}\sum_{m=1}^{M}\mathrm{Rad}_{n_{\text{aux}}}(\mathcal{H}_m|_j)\right) + \max_j \mathrm{Rad}_{n_{\text{aux}}}(\mathcal{G}|_j)\right]\right)
\end{aligned}
$$

Proof. We define two function classes

$$
\mathcal{Q}_{\mathcal{H}} := \{(x,y) \mapsto \psi\left(h(x)_y\right) : h \in \mathcal{H}\} \quad \text{and} \quad \mathcal{Q}_{\mathcal{G}} := \{(x,y) \mapsto \psi\left(g(x)_y\right) : g \in \mathcal{G}\},
$$

and use the fact that:

$$
\frac{1}{n_{\text{aux}}}\sum_{i=1}^{n_{\text{aux}}} \left\|\psi\left(g\left(x_i\right)\right) - \psi\left(h\left(x_i\right)\right)\right\|_1 = \frac{1}{n_{\text{aux}}}\sum_{i=1}^{n_{\text{aux}}} \left\|1 - g\left(x_i\right) - 1 + h\left(x_i\right)\right\|_1 = \Phi_{\mu_{\text{aux}},n_{\text{aux}}}(h_1,\ldots,h_M; g).
$$

We use $\mathcal{Q}_{\mathcal{H}}$ and $\mathcal{Q}_{\mathcal{G}}$ in Lemma 6, and use Lemma 4 and Lemma 5 to estimate $\mathrm{Rad}_n\left(\mathcal{Q}_{\mathcal{H}}\right)$ and $\mathrm{Rad}_n\left(\mathcal{Q}_{\mathcal{G}}\right)$, with probability $1-\delta$, yielding

$$
\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mu}[\psi(g(x))_y] \\
&\leq \mathbb{E}_{(x,y)\sim\mu}[\psi(h(x))_y] + \frac{1}{n_{\mathsf{aux}}}\sum_{i=1}^{n_{\mathsf{aux}}} \min\left\{1, \|\psi(g(x_i)) - \psi(h(x_i))\|_1\right\} + 2\mathbb{TV}(\mu_{\mathsf{aux},\mu}) + 3\sqrt{\frac{\log(2/\delta)}{2n_{\mathsf{aux}}}} \\
&\quad + 2\sum_{y'=1}^{k}\left(\mathrm{Rad}_{n_{\mathsf{aux}}}\left(\{x \mapsto \psi(h(x))_{y'} : h \in \mathcal{H}\}\right) + \mathrm{Rad}_{n_{\mathsf{aux}}}\left(\{x \mapsto \psi(g(x))_{y'} : g \in \mathcal{G}\}\right)\right) \\
&\leq \mathbb{E}_{(x,y)\sim\mu}[1 - h(x)_y)] + \Phi_{\mu_{\mathsf{aux}}, n_{\mathsf{aux}}}(h_1, \ldots, h_M; g) + 2\mathbb{TV}(\mu_{\mathsf{aux}}, \mu) + 3\sqrt{\frac{\log(2/\delta)}{2n_{\mathsf{aux}}}} \\
&\quad + \mathcal{O}\left(k^{3/2}\left[\max_j \mathrm{Rad}_{n_{\mathsf{aux}}}(\mathcal{H}|_j) + \max_j \mathrm{Rad}_{n_{\mathsf{aux}}}(\mathcal{G}|_j)\right]\right) \qquad \text{(Due to Equation (6) and Lemma 4)} \\
&= \mathbb{E}_{(x,y)\sim\mu}[1 - h(x)_y] + \Phi_{\mu_{\mathsf{aux}}, n_{\mathsf{aux}}}(h_1, \ldots, h_M; g) + 2\mathbb{TV}(\mu_{\mathsf{aux}}, \mu) + 3\sqrt{\frac{\log(2/\delta)}{2n_{\mathsf{aux}}}} \\
&\quad + \mathcal{O}\left(k^{3/2}\left[\max_j \left(\frac{1}{M}\sum_{m=1}^{M}\mathrm{Rad}_{n_{\mathsf{aux}}}(\mathcal{H}_m|_j)\right) + \max_j \mathrm{Rad}_{n_{\mathsf{aux}}}(\mathcal{G}|_j)\right]\right) \qquad \text{(Due to Lemma 5)}
\end{aligned}
$$

□

To show our generalization bounds in Theorem 1, it remains to bound $\mathbb{E}_{(x,y)\sim\mu}[1 - h(x)_y]$ in Lemma 7.

Lemma 8. *Let classes of bounded functions $\mathcal{H}_m$ be given with $h_m \in \mathcal{H}_m : \mathcal{X} \to [0,1]^k$, $\forall m \in [M]$, and d_m be the VC dimension of $\mathcal{H}_m$. Then with probability at least $1 - \delta$ over the draw of $\mathcal{D}' = \{(x_i, y_i)\}_{i=1}^n$ from distribution μ, and $\mathcal{D}'_m$ from distribution μ_m with size n , for every $h_m \in \mathcal{H}_m, \forall m \in [M]$,*

$$
\begin{aligned}
\mathbb{E}_{(x,y)\sim\mu}\left[1 - h(x)_y\right] \leq & \frac{1}{M}\sum_{m=1}^{M}\Bigg(\mathbb{E}_{(x,y)\sim\mu_m}\left[1 - h_m(x)_y\right] + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}'_m, \mathcal{D}') \\
& + \lambda_m + 4\sqrt{\frac{2d_m \log(2n) + \log(2M/\delta)}{n}}\Bigg)
\end{aligned}
$$

where $\lambda_m = \varepsilon_{\mu_m}(h^*) + \varepsilon_\mu(h^*)$ and $h^* := \arg\min_{h\in\mathcal{H}} \varepsilon_{\mu_m}(h) + \varepsilon_\mu(h)$.

Proof. Since the predictions from different local models h_m are independent, we can expand $h(x)$ as below:

$$
\mathbb{E}_{(x,y)\sim\mu}\left[1 - h(x)_y\right] = \mathbb{E}_{(x,y)\sim\mu}\left[1 - \left(\frac{1}{M}\sum_{m=1}^{M} h_m(x)_y\right)\right] = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{(x,y)\sim\mu}\left[1 - h_m(x)_y\right]
$$

We apply Lemma 2 for the target distribution μ and each local distribution μ_m. Concretely, with probability $1 - \delta/M$,

$$
\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mu}\left[1-h_m(x)_y\right] \\
=&\mathbb{E}_{(x,y)\sim\mu}|h_m(x)_y-h^*_\mu(x)_y| && \text{(use the fact of labeling function that } h^*_\mu(x)_y=1,(x,y)\sim\mu) \\
=&\varepsilon_\mu(h_m) && \text{(use the labeling function as in Definition 1)} \\
\leq&\varepsilon_{\mu_m}(h_m)+\frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}'_m,\mathcal{D}')+4\sqrt{\frac{2d_m\log(2n)+\log(2M/\delta)}{n}}+\lambda_m \\
=&\mathbb{E}_{(x,y)\sim\mu_m}|h_m(x)_y-h^*_{\mu_m}(x)_y|+\frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}'_m,\mathcal{D}')+4\sqrt{\frac{2d_m\log(2n)+\log(2M/\delta)}{n}}+\lambda_m \\
=&\mathbb{E}_{(x,y)\sim\mu_m}\left[1-h_m(x)_y\right]+\frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}'_m,\mathcal{D}')+4\sqrt{\frac{2d_m\log(2n)+\log(2M/\delta)}{n}}+\lambda_m \\
& && \text{(use the fact of labeling functions that } h^*_{\mu_m}(x)_y=1,(x,y)\sim\mu_m)
\end{aligned}
$$

where $\lambda_m=\varepsilon_{\mu_m}(h^*)+\varepsilon_\mu(h^*)$ and $h^*:=\arg\min_{h\in\mathcal{H}}\varepsilon_{\mu_m}(h)+\varepsilon_\mu(h)$.

Combing all $m\in[M]$ together, with with probability $1-\delta$, we have

$$
\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mu}\left[1-h(x)_y\right] \\
=&\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{(x,y)\sim\mu}\left[1-h_m(x)_y\right] \\
\leq&\frac{1}{M}\sum_{m=1}^{M}\left(\mathbb{E}_{(x,y)\sim\mu_m}\left[1-h_m(x)_y\right]+\frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}'_m,\mathcal{D}')+\lambda_m+4\sqrt{\frac{2d_m\log(2n)+\log(2M/\delta)}{n}}\right)
\end{aligned}
$$

□

Lemma 9. *With probability at least $1-\delta$, we have w.r.t the draw of $\mathbb{D}_m\sim\mu_m$ with $|\mathbb{D}_m|=n$ that*

$$
\mathbb{E}_{(x,y)\sim\mu_m}\left[1-h_m(x)_y\right]\leq \mathtt{ERR}(\mathbb{D}_m,h_m)+2\,\mathrm{Rad}_n(\mathcal{H}_m)+3\sqrt{\frac{\log(2/\delta)}{2n}} \tag{10}
$$

where $\mathtt{ERR}(\mathbb{D}_m,h_m)=\frac{1}{n}\sum_{j=1}^{n}\left[1-h_m(x_{m,j})_{y_{m,j}}\right]$.

Proof. The proofs directly follow Lemma 1 with $b=1,a=0$. □

Given Lemma 7 and Lemma 8 with at least $1-\delta/3$ probability for each event, and Lemma 9 with at least $1-\delta/3M$ probability for each local model $m\in[M]$, we can bound $\mathbb{E}_{x,y}\psi(g(x))_y$ in Equation (9), which proves the main result in Theorem 1.

C.4. Proof for Generalization Bounds of Personalized Models in Theorem 2

Overview For the generalization bounds of the personalized models, we will upper bound error probabilities of p_m with the expected prediction distances between the global model and personalized model on μ as well as errors of the global model on μ.

Main Analysis The proofs for Theorem 2 are similar to Lemma 6 and Lemma 7. We first introduce Lemma 10 as below.

Lemma 10. *Let classes of bounded functions $\mathcal{P}_m$ and $\mathcal{G}$ be given with $p_m\in\mathcal{P}_m:\mathcal{X}\to[0,1]^k$ and $g\in\mathcal{G}:\mathcal{X}\to[0,1]^k$. Suppose $\{x_i\}_{i=1}^n$ is sampled from a distribution μ. For every $p_m\in\mathcal{P}_m$ and every $g\in\mathcal{G}$, with probability at least $1-\delta$,*

$$
\begin{aligned}
\mathbb{E}_{(x,y)\sim\mu}p_m(x)_y\leq\mathbb{E}_{(x,y)\sim\mu}g(x)_y+\frac{1}{n}\sum_{i=1}^{n}\min\{1,\|p_m(x_i)-g(x_i)\|_1\}+3\sqrt{\frac{\log(2/\delta)}{2n}} \\
+2\sum_{y'=1}^{k}\left(\mathrm{Rad}_n\left(\{x\mapsto p_m(x)_{y'}:p_m\in\mathcal{P}_m\}\right)+\mathrm{Rad}_n\left(\{x\mapsto g(x)_{y'}:g\in\mathcal{G}\}\right)\right)
\end{aligned}
$$

Proof. To start, for any $p_m \in \mathcal{P}_m, g \in \mathcal{G}$, write

$$\mathbb{E}_{x,y} p_m(x)_y = \mathbb{E}_{x,y}(p_m(x) - g(x))_y + \mathbb{E}_{x,y} g(x)_y$$

For the first term, since $p_m : \mathcal{X} \to [0,1]^k$ and $g : \mathcal{X} \to [0,1]^k$, by Holder's inequality

$$\mathbb{E}_{x,y}(p_m(x) - g(x))_y = \int \min\left\{1, (p_m(x) - g(x))_y\right\} \mathrm{d}\mu(x,y) \leq \int \min\left\{1, \|p_m(x) - g(x)\|_1\right\} \mathrm{d}\mu_{\mathcal{X}}(x)$$

Once again invoking standard Rademacher complexity arguments Lemma 1, with probability at least $1-\delta$, every mapping $x \mapsto \min\left\{1, \|p_m(x) - g(x)\|_1\right\}$ where $p_m \in \mathcal{P}_m$ and $g \in \mathcal{G}$ satisfies

$$\begin{aligned}
&\int \min\left\{1, \|p_m(x_i) - g(x_i)\|_1\right\} \mathrm{d}\mu_{\mathcal{X}}(x) \\
&\leq \int \min\{1, \|p_m(x_i) - g(x_i)\|_1\} \mathrm{d}\mu(x) \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \min\left\{1, \|p_m(x_i) - g(x_i)\|_1\right\} + 3\sqrt{\frac{\log(2/\delta)}{2n}} \\
&+ 2\,\mathrm{Rad}_n\left(\left\{x \mapsto \min\left\{1, \|p_m(x_i) - g(x_i)\|_1\right\} : p_m \in \mathcal{P}_m, g \in \mathcal{G}\right\}\right)
\end{aligned}$$

For the final Rademacher complexity estimate, we follow the proofs in our previous Lemma 6 and have

$$\begin{aligned}
&\mathrm{Rad}_n\left(\left\{x \mapsto \min\left\{1, \|p_m(x) - g(x)\|_1\right\} : p_m \in \mathcal{P}_m, g \in \mathcal{G}\right\}\right) \\
&\leq \sum_{y'=1}^{k} \mathrm{Rad}_n\left(\left\{x \mapsto |p_m(x) - g(x)|_{y'} : p_m \in \mathcal{P}_m, g \in \mathcal{G}\right\}\right).
\end{aligned}$$

Also following the proof steps in our Lemma 6, we have for every $y' \in [k]$

$$\begin{aligned}
&\mathrm{Rad}_n\left(\left\{x \mapsto |p_m(x) - g(x)|_{y'} : p_m \in \mathcal{P}_m, g \in \mathcal{G}\right\}\right) \\
&\leq \mathrm{Rad}_n\left(\left\{x \mapsto p_m(x)_{y'} : p_m \in \mathcal{P}_m\right\}\right) + \mathrm{Rad}_n\left(\left\{x \mapsto g(x)_{y'} : g \in \mathcal{G}\right\}\right)
\end{aligned}$$

Combining the above results together, we complete the proof. □

Let us recall Theorem 2.

Theorem 2 (Generalization bound of PERADA personalized model)**.** *With probability at least* $1-\delta$*, for every* $p_m \in \mathcal{P}_m, \forall m \in [M]$*, and for every* $g \in \mathcal{G}$*, we have* $\mathrm{Pr}_{(x,y)\sim\mu}\left[\underset{y'}{\arg\max} p_m(x)_{y'} \neq y\right] \leq 2\mathbb{E}_{(x,y)\sim\mu}(1 - g(x)_y) + 2\frac{1}{n}\sum_{i=1}^{n} \min\left\{1, \|p_m(x) - g(x)\|_1\right\} + 6\sqrt{\frac{\log(2/\delta)}{2n}} + \mathcal{O}\left(k^{3/2}\left[\max_j \mathrm{Rad}_n(\mathcal{P}|_j) + \max_j \mathrm{Rad}_n(\mathcal{G}|_j)\right]\right)$.

Then we prove Theorem 2 as below:

Proof for Theorem 2. Following the proofs in our previous Lemma 7, we define two function classes

$$\mathcal{Q}_{\mathcal{P}_m} := \left\{(x,y) \mapsto \psi\left(p_m(x)_y\right) : p_m \in \mathcal{P}_m\right\} \quad \text{and} \quad \mathcal{Q}_{\mathcal{G}} := \left\{(x,y) \mapsto \psi\left(g(x)_y\right) : g \in \mathcal{G}\right\},$$

and use the fact that:

$$\frac{1}{n}\sum_{i=1}^{n} \left\|\psi\left(p_m\left(x_i\right)\right) - \psi\left(g\left(x_i\right)\right)\right\|_1 = \frac{1}{n}\sum_{i=1}^{n} \left\|1 - p_m\left(x_i\right) - 1 + g\left(x_i\right)\right\|_1 = \frac{1}{n}\sum_{i=1}^{n} \left\|p_m\left(x_i\right) - g\left(x_i\right)\right\|_1$$

We use $\mathcal{Q}_{\mathcal{P}_m}$ and $\mathcal{Q}_{\mathcal{G}}$ in Lemma 10, and use Lemma 4 and Lemma 5 to estimate $\mathrm{Rad}_n\left(\mathcal{Q}_{\mathcal{P}_m}\right)$ and $\mathrm{Rad}_n\left(\mathcal{Q}_{\mathcal{G}}\right)$, with probability $1-\delta$, yielding

$$
\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mu}[1 - p_m(x)_y] = \mathbb{E}_{(x,y)\sim\mu}[\psi(p_m(x))_y] \\
&\leq \mathbb{E}_{(x,y)\sim\mu}[\psi(g(x))_y] + \frac{1}{n}\sum_{i=1}^{n} \min\left\{1, \|\psi(p_m(x_i)) - \psi(g(x_i))\|_1\right\} + 3\sqrt{\frac{\log(2/\delta)}{2n}} \\
&\quad + 2\sum_{y'=1}^{k} \left(\mathrm{Rad}_n\left(\{x \mapsto \psi(p_m(x))_{y'} : p_m \in \mathcal{P}_m\}\right) + \mathrm{Rad}_n\left(\{x \mapsto \psi(g(x))_{y'} : g \in \mathcal{G}\}\right)\right) \\
&\leq \mathbb{E}_{(x,y)\sim\mu}[1 - g(x)_y] + \frac{1}{n}\sum_{i=1}^{n} \min\left\{1, \|p_m(x_i) - g(x_i)\|_1\right\} + 3\sqrt{\frac{\log(2/\delta)}{2n}} \\
&\quad + \mathcal{O}\left(k^{3/2}\left[\max_j \mathrm{Rad}_n(\mathcal{P}_m|_j) + \max_j \mathrm{Rad}_n(\mathcal{G}|_j)\right]\right) \qquad \text{(Due to Lemma 4)}
\end{aligned}
$$

Finally, we use Lemma 3 to show that

$$
\mathbb{E}_{x,y}\left(1 - p_m(x)_y\right) \geq \frac{1}{2}\mathbb{E}_{x,y}\left[\mathbb{1}\left[\arg\max_{y'} p_m(x)_{y'} \neq y\right]\right] = \frac{1}{2}\,\mathrm{Pr}_{x,y}\left[\arg\max_{y'} p_m(x)_{y'} \neq y\right]
$$

Combining all results together, with probability at least $1 - \delta$, we have,

$$
\begin{aligned}
\mathrm{Pr}_{x,y}\left[\arg\max_{y'} p_m(x)_{y'} \neq y\right] &\leq 2\mathbb{E}_{(x,y)\sim\mu}[1 - g(x)_y] + 2\frac{1}{n}\sum_{i=1}^{n} \min\left\{1, \|p_m(x_i) - g(x_i)\|_1\right\} + 6\sqrt{\frac{\log(2/\delta)}{2n}} \\
&\quad + \mathcal{O}\left(k^{3/2}\left[\max_j \mathrm{Rad}_n(\mathcal{P}_m|_j) + \max_j \mathrm{Rad}_n(\mathcal{G}|_j)\right]\right)
\end{aligned}
$$

This completes the proof.

□

D. Convergence Analysis

In this section, we present the discussions and analysis for our convergence guarantees. The outline of this section is as follows:

- Appendix D.1 provides more discussions and additional convergence results.
- Appendix D.2 provides the proofs for the global model convergence guarantee in Theorem 3.
- Appendix D.3 provides the proofs for the personalized model convergence guarantee in Theorem 4.

D.1. Additional Discussions and Theoretical Results

Discussions on distillation gradient For simplicity, we denote $\overline{f(\theta, x)} = \frac{1}{M}\sum_{m=1}^{M} f(\theta_m, x)$. The closed-form expression of $\nabla_w \mathcal{R}$ can be expressed as:

$$\begin{aligned}
&\left\|\nabla_w \mathcal{R}\left(\{\theta_1, \ldots, \theta_M\}, w\right)\right\| \\
&= \left\|\mathbb{E}_{x\sim\mu_{\text{aux}}} \sum_{i=1}^{k} \nabla_w \left[\sigma(\overline{f(\theta, x)})_i \ln\left(\frac{\sigma(\overline{f(\theta, x)})_i}{\sigma(f(w, x))_i}\right)\right]\right\| && \text{(KL divergence loss)} \\
&= \left\|\mathbb{E}_{x\sim\mu_{\text{aux}}} \sum_{i=1}^{k} -\frac{\sigma(\overline{f(\theta, x)})_i}{\sigma(f(w, x))_i} \nabla_w \sigma(f(w, x))_i\right\| \\
&= \left\|\mathbb{E}_{x\sim\mu_{\text{aux}}} \sum_{i=1}^{k} \frac{\sigma(\overline{f(\theta, x)})_i}{\sigma(f(w, x))_i} \nabla_w \sigma(f(w, x))_i\right\| && (11)
\end{aligned}$$

where k is the number of classes and i denotes the i-th class. Here we note that when the averaged logits from local models are qual to the logits of global model, i.e., $\sigma(\overline{f(\theta, x)})_i = \sigma(f(w, x))_i$

$$\left\|\nabla_w \mathcal{R}\left(\{\theta_1, \ldots, \theta_M\}, w\right)\right\| = \left\|\mathbb{E}_{x\sim\mu_{\text{aux}}} \sum_{i=1}^{k} \nabla_w \sigma(f(w, x))_i\right\| = 0 \tag{12}$$

because $\sum_{i=1}^{k} \sigma(f(w, x))_i = 1$ (which leads to $\nabla_w \sum_{i=1}^{k} \sigma(f(w, x))_i = 0$) . Therefore, the norm of distillation gradient can be small when the averaged logits from local models are close to the logits of global model.

Discussions for Assumptions. Assumption 1 on Lipschitz smooth and Assumption 2 on the bounded variance for gradients due to stochastic sampling noise are standard for smooth and non-convex optimization. Assumption 3 quantifies the diversity of FL clients' data distribution, which is widely used in FL optimization [12, 27, 35, 46, 52]. We follow [12, 46, 52] to assume bounded gradient for non-convex FL optimization in Assumption 4.

Convergence of PERADA personalized models.

Theorem 4 (Convergence of PERADA personalized model)**.** *When $\eta_p = \frac{1}{(L+\lambda)\sqrt{T}}$, $\eta_l = \frac{1}{EL\sqrt{T}}$, $\eta_g = \frac{1}{L_R RT}$, then the algorithm satisfies:*

$$\frac{1}{TS}\sum_{t=0}^{T-1}\sum_{s=0}^{S-1} \mathbb{E}\|\nabla_v P_m(v_m^{t,s}, w^t)\|^2 \le \mathcal{O}\Big(\frac{(L+\lambda)\Delta_{P_m} + \phi_2}{\sqrt{T}S} + \frac{G_P(L+\lambda)(L\Delta_{\mathcal{L}} + \psi_1)^{1/2}}{T^{1/4}L\sqrt{E}S} + \frac{G_P(L+\lambda)\sqrt{\psi_2}}{T^{3/4}L_R ES} + \frac{G_P(L+\lambda)\sigma}{LES}\Big)$$

where $\Delta_{P_m} = P_m(v_m^0, w_m^0) - P_m(v_m^t, w^t)$, $\phi_1 = 64(3\bar{\gamma} + \frac{2\sigma^2}{E})$, $\phi_2 = S\sigma^2 + \frac{\sqrt{\phi_1}G_P(L+\lambda)}{LE} + \sqrt{\psi_2}G_P(L+\lambda) + \frac{G_P\bar{\gamma}(L+\lambda)}{L\sqrt{E}}$. ψ_1, ψ_2 are defined the same as in Theorem 3.

Remark 4. (1) **Local steps**: a larger local step S can reduce number of rounds T for convergence. (2) **Connection to global model**: The terms associated with $\bar{\gamma}, \psi_1, \psi_2$ are related to the convergence rate of the global model, which is indicated in Theorem 3. For example, a large E can also reduce the number of communication rounds T for the personalized model to convergence. We obtain a convergence rate of $\mathcal{O}(1/T^{1/4})$ for personalized models. It is worth noting that previous studies have shown that in strongly convex settings, personalized models converge at the same rate as the global model [33]. However, in strongly convex settings, the minimizers are ensured to be unique, which can simplify the establishment of connections between global and personalized models by considering their distances to the corresponding minimizers. Here, we present the results in the more general non-convex setting and additionally analyze the effect of the global model's ensemble distillation on personalized models.

D.2. Proofs for the Global Model Convergence Guarantee in Theorem 3

Additional notations Recall the parameter-averaged model is $\bar{\theta}^{t+1} = \frac{1}{M}\sum_{m=1}^{M} \theta_m^{t+1}$, which is used to initialize the server global model at round t before the KD training. Let

$$\bar{\eta}_g = \eta_g R, \quad \bar{\eta}_l = \eta_l E \tag{13}$$

Based on the update rules, we define g^t and g_m^t as below, which capture the update of global model during server training, and the update of local model during client training, respectively.

$$w^{t+1} = \bar{\theta}^{t+1} - \bar{\eta}_g g^t, \quad \theta_m^{t+1} = w^t - \bar{\eta}_l g_m^t \tag{14}$$

That is:

$$\begin{aligned} g^t &:= -\frac{1}{\eta_g R}(w^{t+1} - \bar{\theta}^{t+1}) = \frac{1}{R}\sum_{r=0}^{R-1} \widetilde{\nabla}_w \mathcal{R}(\{\theta_m^{t+1}\}, w^{t,r}), \\ g_m^t &:= -\frac{1}{\eta_l E}(\theta_m^{t+1} - w^t) = \frac{1}{E}\sum_{e=0}^{E-1} \widetilde{\nabla} \mathcal{L}_m(\theta_m^{t,e}) \end{aligned} \tag{15}$$

According to server update rule $w^{t+1} - w^t = \bar{\theta}^{t+1} - w^t - \bar{\eta}_g g^t$. Note that $\bar{\theta}^{t+1} - w^t = \frac{1}{M}\sum_{m=1}^{M} \theta_m^{t+1} - w^t = -\frac{1}{M}\sum_{m=1}^{M} \bar{\eta}_l g_m^t$ based on Equation (15). Then we define,

$$\delta_w^t := \frac{1}{M}\sum_{m=1}^{M} g_m^t + \frac{\bar{\eta}_g}{\bar{\eta}_l} g^t, \quad \text{which indicates } w^{t+1} - w^t = -\bar{\eta}_l \delta_w^t \tag{16}$$

According to client update rule $\theta_m^{t+1} - \theta_m^t = (w^t - \bar{\eta}_l g_m^t) - (w^{t-1} - \bar{\eta}_l g_m^{t-1})$. Note that $w^t - w^{t-1} = -\bar{\eta}_l \frac{1}{M}\sum_{m=1}^{M} g_m^t - \bar{\eta}_g g^t$ based on Equation (15). Then we define,

$$\delta_{\theta_m}^t := \frac{\bar{\eta}_g}{\bar{\eta}_l} g^{t-1} + \frac{1}{M}\sum_{i=1}^{M} g_i^{t-1} - g_m^{t-1} + g_m^t, \quad \text{which indicates } \theta_m^{t+1} - \theta_m^t = -\bar{\eta}_l \delta_{\theta_m}^t \tag{17}$$

In our analysis, we define one virtual sequence $\bar{w}^{t,e}$, motivated by [35],

$$\bar{w}^{t,e} = \frac{1}{M}\sum_{m=1}^{M} \theta_m^{t,e} \tag{18}$$

In particular, $\bar{w}^{t+1,0} = w^t$ and $\bar{w}^{t+1,E-1} = \bar{\theta}^{t+1}$.

Proof Outline Recall the generic FL objective, which is to minimize the average loss measured on all clients' data:

$$\mathcal{L}(w) := \frac{1}{M}\sum_{m=1}^{M} \mathcal{L}_m(w) \tag{19}$$

The goal is to bound the gradients of the global model w.r.t the $\mathcal{L}(w)$, which is used to show that the trained models can converge to the stationary points:

$$\sum_{t=0}^{T-1}\sum_{e=0}^{E-1} \frac{1}{ET}\mathbb{E}\|\nabla \mathcal{L}(\bar{w}^{t,e})\|^2 \tag{20}$$

Challenges The challenges of convergence analysis include: (1) Bi-level optimization between server distillation for w^t and client training for $\{\theta_m^t\}$, which incorporates two objectives (i.e., minimizing distillation loss and local loss respectively), as shown in Equation (15). (2) Mutual initializations. At each round, the global model is initialized by averaged local models before distillation, and local models are initialized by the global model before local training. Such mutual initializations intervene in the model updating trajectories of w^t and $\{\theta_m^t\}$ w.r.t their training objective. In particular, the server optimization of w will be influenced by the drift of client optimization of θ_m, as shown in Equation (16) (i.e., additional deviation with the term $\frac{1}{M}\sum_{m=1}^{M} g_m^t$). Moreover, client optimization is also influenced by the drift of server optimization, as shown in Equation (17) (i.e., additional deviation with the terms $\frac{\bar{\eta}_g}{\bar{\eta}_l} g^{t-1} + \frac{1}{M}\sum_{i=1}^{M} g_i^{t-1} - g_m^{t-1}$).

To overcome the aforementioned challenges, we regard $\{\theta_m^t\}$ as the intermediate models to update w^{t+1}, and quantify the effects of local client updates and server distillation updates on reducing $\mathcal{L}(w^{t+1})$.

Supporting lemmas Before we start, we introduce a useful existing result by Jensen's inequality in Lemma 11:

Lemma 11 (Jensen's inequality)**.** *For any vector $x_i \in \mathbb{R}^d, i = 1, \ldots, M$, by Jensen's inequality, we have*

$$\left\|\sum_{i=1}^{M} x_i\right\|^2 \leq M \sum_{i=1}^{M} \|x_i\|^2 \tag{21}$$

We also introduce the following supporting lemmas:

Lemma 12 (Bounded local client drift error)**.** *If $\bar{\eta}_l \leq \frac{1}{2L} \Leftrightarrow \eta_l \leq \frac{1}{2LE}$, we have*

$$\mathbb{E}\left[\left\|\widetilde{\nabla}\mathcal{L}_m(\theta_m^{t,e}) - \nabla\mathcal{L}_m(w^t)\right\|^2\right] \leq 2\sigma^2 + 16L^2\bar{\eta}_l^2\left(3\mathbb{E}\left[\left\|\nabla\mathcal{L}_m(w^t)\right\|^2\right] + \frac{2\sigma^2}{E}\right). \tag{22}$$

Moreover, the averaged drift error over E local steps and M clients is:

$$\frac{1}{ME}\sum_{m,e}^{M,E}\mathbb{E}\left[\left\|\widetilde{\nabla}\mathcal{L}_m(\theta_m^{t,e}) - \nabla\mathcal{L}_m(w^t)\right\|^2\right] \leq 2\sigma^2 + 16L^2\bar{\eta}_l^2\left(3\bar{\gamma} + 3\mathbb{E}\left[\left\|\nabla\mathcal{L}(w^t)\right\|^2\right] + \frac{2\sigma^2}{E}\right). \tag{23}$$

Proof.

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{\nabla}\mathcal{L}_m(\theta_m^{t,e}) - \nabla\mathcal{L}_m(w^t)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\widetilde{\nabla}\mathcal{L}_m(\theta_m^{t,e}) - \nabla\mathcal{L}_m(\theta_m^{t,e}) + \nabla\mathcal{L}_m(\theta_m^{t,e}) - \nabla\mathcal{L}_m(w^t)\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\widetilde{\nabla}\mathcal{L}_m(\theta_m^{t,e}) - \nabla\mathcal{L}_m(\theta_m^{t,e})\right\|^2\right] + 2\mathbb{E}\left[\left\|\nabla\mathcal{L}_m(\theta_m^{t,e}) - \nabla\mathcal{L}_m(w^t)\right\|^2\right] \\
&\leq 2\sigma^2 + 2\mathbb{E}\left[\left\|\nabla\mathcal{L}_m(\theta_m^{t,e}) - \nabla\mathcal{L}_m(w^t)\right\|^2\right] && \text{(Assumption 2)} \\
&\leq 2\sigma^2 + 2L^2\mathbb{E}\left[\left\|\theta_m^{t,e} - w^t\right\|^2\right] && \text{(Assumption 1)}
\end{aligned}$$

If $\bar{\eta}_l \leq \frac{1}{2L} \Leftrightarrow \eta_l \leq \frac{1}{2LE}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\theta_m^{t,e} - w^t\right\|^2\right] = \mathbb{E}\left[\left\|\theta_m^{t,e-1} - w^t - \eta_l \widetilde{\nabla}\mathcal{L}_m(\theta_m^{t,e-1})\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\theta_m^{t,e-1} - w^t - \eta_l \nabla\mathcal{L}_m(w^t) + \eta_l \nabla\mathcal{L}_m(w^t) - \eta_l \widetilde{\nabla}\mathcal{L}_m(\theta_m^{t,e-1})\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\theta_m^{t,e-1} - w^t - \eta_l \nabla\mathcal{L}_m(w^t)\right\|^2\right] + 2\eta_l^2 \mathbb{E}\left[\left\|\nabla\mathcal{L}_m(w^t) - \widetilde{\nabla}\mathcal{L}_m(\theta_m^{t,e-1})\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\theta_m^{t,e-1} - w^t\right\|^2\right] + 2\mathbb{E}\left[\left\|\eta_l \nabla\mathcal{L}_m(w^t)\right\|^2\right] - 2\mathbb{E}\left[\langle \theta_m^{t,e-1} - w^t, \eta_l \nabla\mathcal{L}_m(w^t)\rangle\right] \\
&\quad + 2\eta_l^2 \mathbb{E}\left[\left\|\nabla\mathcal{L}_m(w^t) - \nabla\mathcal{L}_m(\theta_m^{t,e-1}) + \nabla\mathcal{L}_m(\theta_m^{t,e-1}) - \widetilde{\nabla}\mathcal{L}_m(\theta_m^{t,e-1})\right\|^2\right] \\
&\leq 2\left(1 + \frac{1}{2E}\right)\mathbb{E}\left[\left\|\theta_m^{t,e-1} - w^t\right\|^2\right] + 2\eta_l^2\left(1 + 2E\right)\mathbb{E}\left[\left\|\nabla\mathcal{L}_m(w^t)\right\|^2\right] \\
&\quad + 4\eta_l^2 L^2 \mathbb{E}\left[\left\|\theta_m^{t,e-1} - w^t\right\|^2\right] + 4\eta_l^2\sigma^2 \\
&= 2\left(1 + \frac{1}{2E} + 2\eta_l^2 L^2\right)\mathbb{E}\left[\left\|\theta_m^{t,e-1} - w^t\right\|^2\right] + 2\eta_l^2\left(1 + 2E\right)\mathbb{E}\left[\left\|\nabla\mathcal{L}_m(w^t)\right\|^2\right] + 4\eta_l^2\sigma^2 \\
&\overset{(a)}{\leq} 2\left(1 + \frac{1}{E}\right)\mathbb{E}\left[\left\|\theta_m^{t,e-1} - w^t\right\|^2\right] + \frac{6\bar{\eta}_l^2}{E}\mathbb{E}\left[\left\|\nabla\mathcal{L}_m(w^t)\right\|^2\right] + \frac{4\bar{\eta}_l^2\sigma^2}{E^2} \\
&\leq 2\sum_{e=0}^{E-1}\left(1 + \frac{1}{E}\right)^e\left(\frac{6\bar{\eta}_l^2}{E}\mathbb{E}\left[\left\|\nabla\mathcal{L}_m(w^t)\right\|^2\right] + \frac{4\bar{\eta}_l^2\sigma^2}{E^2}\right) \\
&\overset{(b)}{\leq} 8\bar{\eta}_l^2\left(3\mathbb{E}\left[\left\|\nabla\mathcal{L}_m(w^t)\right\|^2\right] + \frac{2\sigma^2}{E}\right)
\end{aligned}$$

Here (a) is because: $\eta_l = \frac{\bar{\eta}_l}{E}$ and when $\bar{\eta}_l^2 \leq \frac{1}{4L^2}$, we have $2\eta_l^2 L^2 = \frac{2\bar{\eta}_l^2 L^2}{E^2} \leq \frac{1}{2E^2} \leq \frac{1}{2E}$ for all $E \geq 1$. Moreover, $2\eta_l^2\left(1 + 2E\right) = 2\left(1 + 2E\right)\frac{\bar{\eta}_l^2}{E^2} \leq \frac{6\bar{\eta}_l^2}{E}$ because $\frac{1+2E}{E} \leq 3$ for $E \geq 1$.

(b) is because: $\sum_{e=0}^{E-1}(1 + 1/E)^e = \frac{(1+1/E)^E - 1}{1/E} \leq \frac{\mathrm{e}-1}{1/E} \leq 2E$ by using the fact $\sum_{i=0}^{n-1} x^i = \frac{x^n - 1}{x-1}$ and $\left(1 + \frac{x}{n}\right)^n \leq e^x$ for any $x \in \mathbb{R}, n \in \mathbb{N}$.

Combining the above results together, we have

$$\mathbb{E}\left[\left\|\widetilde{\nabla}\mathcal{L}_m(\theta_m^{t,e}) - \nabla\mathcal{L}_m(w^t)\right\|^2\right] \leq 2\sigma^2 + 16L^2\bar{\eta}_l^2\left(3\mathbb{E}\left[\left\|\nabla\mathcal{L}_m(w^t)\right\|^2\right] + \frac{2\sigma^2}{E}\right)$$

By the expectation $E[\|X\|^2] = E[\|X - E[X]\|^2] + E[\|X\|]^2$, we have the averaged error over M clients:

$$\begin{aligned}
\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\left\|\nabla\mathcal{L}_m(w^t)\right\|^2\right] &= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\left\|\nabla\mathcal{L}_m(w^t) - \nabla\mathcal{L}(w^t)\right\|^2\right] + \mathbb{E}\left[\left\|\nabla\mathcal{L}(w^t)\right\|^2\right] \\
&\leq \bar{\gamma} + \mathbb{E}\left[\left\|\nabla\mathcal{L}(w^t)\right\|^2\right] \qquad \text{(Assumption 3)}
\end{aligned}$$

Moreover, the averaged error over M clients and E local steps is:

$$\frac{1}{ME}\sum_{m,e}^{M,E}\mathbb{E}\left[\left\|\nabla\mathcal{L}_m(w^t)\right\|^2\right] \leq 2\sigma^2 + 16L^2\bar{\eta}_l^2\left(3\bar{\gamma} + 3\mathbb{E}\left[\left\|\nabla\mathcal{L}(w^t)\right\|^2\right] + \frac{2\sigma^2}{E}\right)$$

Thus, proved. □

Lemma 13 (Bounded distillation drift error)**.** *If $\bar{\eta}_g \leq \frac{1}{2L_R} \Leftrightarrow \eta_g \leq \frac{1}{2RL_R}$, we have*

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_w \mathcal{R}(\{\theta_m^{t+1}\}, w^{t,r}) - \nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^t)\right\|^2\right] \tag{24}$$

$$\leq 2\sigma_R^2 + 16L_R^2\bar{\eta}_g^2\left(3\mathbb{E}\left[\left\|\nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^t)\right\|^2\right] + \frac{2\sigma_R^2}{R}\right). \tag{25}$$

Proof of Lemma 13.

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{\nabla}_w \mathcal{R}(\{\theta_m^{t+1}\}, w^{t,r}) - \nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^t)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\widetilde{\nabla}_w \mathcal{R}(\{\theta_m^{t+1}\}, w^{t,r}) - \nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^{t,r}) + \nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^{t,r}) - \nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^t)\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\widetilde{\nabla}_w \mathcal{R}(\{\theta_m^{t+1}\}, w^{t,r}) - \nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^{t,r})\right\|^2\right] + 2\mathbb{E}\left[\left\|\nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^{t,r}) - \nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^t)\right\|^2\right] \\
&\leq 2\sigma_R^2 + 2L_R^2\mathbb{E}\left[\left\|w^{t,r} - w^t\right\|^2\right] \qquad \text{(Assumption 2, Assumption 1)}
\end{aligned}$$

If $\bar{\eta}_g \leq \frac{1}{2L_R} \Leftrightarrow \eta_g \leq \frac{1}{2RL_R}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|w^{t,r} - w^t\right\|^2\right] = \mathbb{E}\left[\left\|w^{t,r-1} - w^t - \eta_g\widetilde{\nabla}_w \mathcal{R}(\{\theta_m^{t+1}\}, w^{t,r-1})\right\|^2\right] \\
&= \mathbb{E}\left[\left\|w^{t,r-1} - w^t - \eta_g\nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^t) + -\eta_g\nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^t) - \eta_g\widetilde{\nabla}_w \mathcal{R}(\{\theta_m^{t+1}\}, w^{t,r-1})\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|w^{t,r-1} - w^t - \eta_g\nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^t)\right\|^2\right] + 2\mathbb{E}\left[\left\|\eta_g\nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^t) - \eta_g\widetilde{\nabla}_w \mathcal{R}(\{\theta_m^{t+1}\}, w^{t,r-1})\right\|^2\right] \\
&\leq 2\left(1 + \frac{1}{2R}\right)\mathbb{E}\left[\left\|w^{t,r-1} - w^t\right\|^2\right] + 2\eta_g^2\left(1 + 2R\right)\mathbb{E}\left[\left\|\nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^t)\right\|^2\right] \\
&\quad + 4\eta_g^2L_R^2\mathbb{E}\left[\left\|w^{t,r-1} - w^t\right\|^2\right] + 4\eta_g^2\sigma_R^2 \\
&= 2\left(1 + \frac{1}{2R} + 2\eta_g^2L_R^2\right)\mathbb{E}\left[\left\|w^{t,r-1} - w^t\right\|^2\right] + 2\eta_g^2\left(1 + 2R\right)\mathbb{E}\left[\left\|\nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^t)\right\|^2\right] + 4\eta_g^2\sigma_R^2 \\
&\overset{(a)}{\leq} 2\left(1 + \frac{1}{R}\right)\mathbb{E}\left[\left\|w^{t,r-1} - w^t\right\|^2\right] + \frac{6\bar{\eta}_g^2}{R}\mathbb{E}\left[\left\|\nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^t)\right\|^2\right] + \frac{4\bar{\eta}_g^2\sigma_R^2}{R^2} \\
&\leq 2\sum_{r=0}^{R-1}\left(1 + \frac{1}{R}\right)^r\left(\frac{6\bar{\eta}_g^2}{R}\mathbb{E}\left[\left\|\nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^t)\right\|^2\right] + \frac{4\bar{\eta}_g^2\sigma_R^2}{R^2}\right) \\
&\overset{(b)}{\leq} 8\bar{\eta}_g^2\left(3\mathbb{E}\left[\left\|\nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^t)\right\|^2\right] + \frac{2\sigma_R^2}{R}\right)
\end{aligned}$$

Here (a) is because: $\eta_g = \frac{\bar{\eta}_g}{R}$ and when $\bar{\eta}_g^2 \leq \frac{1}{4L_R^2}$, we have $2\eta_g^2L_R^2 = \frac{2\bar{\eta}_g^2L_R^2}{R^2} \leq \frac{1}{2R^2} \leq \frac{1}{2R}$ for all $R \geq 1$. Moreover, $2\eta_g^2\left(1 + 2R\right) = 2\left(1 + 2R\right)\frac{\bar{\eta}_g^2}{R^2} \leq \frac{6\bar{\eta}_g^2}{R}$ because $\frac{1+2R}{R} \leq 3$ for $R \geq 1$.

(b) is because: $\sum_{e=0}^{R-1}(1 + 1/R)^e = \frac{(1+1/R)^R - 1}{1/R} \leq \frac{\mathrm{e}-1}{1/R} \leq 2R$ by using the fact $\sum_{i=0}^{n-1} x^i = \frac{x^n-1}{x-1}$ and $\left(1 + \frac{x}{n}\right)^n \leq e^x$ for any $x \in \mathbb{R}, n \in \mathbb{N}$.

Combining the above results, we have

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_w \mathcal{R}(\{\theta_m^{t+1}\}, w^{t,r}) - \nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^t)\right\|^2\right] \leq 2\sigma_R^2 + 16L_R^2\bar{\eta}_g^2\left(3\mathbb{E}\left[\left\|\nabla_w \mathcal{R}(\{\theta_m^{t+1}\}, w^t)\right\|^2\right] + \frac{2\sigma_R^2}{R}\right)$$

□

Recall Equation (14), we have

$$\bar{\theta}^{t+1} = \frac{1}{M}\sum_{m=1}^{M}\theta_m^{t+1} = \frac{1}{M}(\sum_{m=1}^{M} w^t - \bar{\eta}_l g_m^t). \tag{26}$$

and we define

$$\bar{w}^{t,e} = \frac{1}{M}\sum_{m=1}^{M}\theta_m^{t,e}. \tag{27}$$

Lemma 14.

$$\mathbb{E}[\mathcal{L}(w^{t+1}) - L(\bar{\theta}^{t+1})] \leq \frac{\bar{\eta}_g}{2}(G^2 + \psi_2) + \frac{\bar{\eta}_g^2 L}{2}\psi_2, \tag{28}$$

where $\psi_2 = 4\sigma_R^2 + 32L_R^2\bar{\eta}_g^2(3G_R^2 + \frac{2\sigma_R^2}{R}) + 2G_R^2$.

Proof.

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(w^{t+1}) - L(\bar{\theta}^{t+1})] &\leq \mathbb{E}[\langle \nabla\mathcal{L}(\bar{\theta}^{t+1}), -\bar{\eta}_g g^t\rangle] + \frac{\bar{\eta}_g^2 L}{2}\mathbb{E}\|g^t\|^2 && (29)\\
&\leq \frac{\bar{\eta}_g}{2}\mathbb{E}\|\nabla\mathcal{L}(\bar{\theta}^{t+1})\|^2 + \frac{\bar{\eta}_g}{2}\mathbb{E}\|g^t\|^2 + \frac{\bar{\eta}_g^2 L}{2}\mathbb{E}\|g^t\|^2 && (30)\\
&= \frac{\bar{\eta}_g}{2}\mathbb{E}\|\frac{1}{M}\sum_{m=1}^{M}\nabla\mathcal{L}_m(\bar{\theta}^{t+1})\|^2 + (\frac{\bar{\eta}_g}{2} + \frac{\bar{\eta}_g^2 L}{2})\mathbb{E}\|g^t\|^2. && \text{(Based on } \mathcal{L} = (\theta)\tfrac{1}{M}\textstyle\sum_{m=1}^{M}\nabla\mathcal{L}_m(\theta)\text{)}\\
&\leq \frac{\bar{\eta}_g}{2}\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|\nabla\mathcal{L}_m(\bar{\theta}^{t+1})\|^2 + (\frac{\bar{\eta}_g}{2} + \frac{\bar{\eta}_g^2 L}{2})\mathbb{E}\|g^t\|^2 && \text{(Lemma 11)}\\
&\leq \frac{\bar{\eta}_g}{2}G^2 + (\frac{\bar{\eta}_g}{2} + \frac{\bar{\eta}_g^2 L}{2})\mathbb{E}\|g^t\|^2. && \text{(Assumption 4)}
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}\|g^t\|^2 &= \mathbb{E}\|\frac{1}{R}\sum_{r=0}^{R-1}\tilde{\nabla}_W R(\{\theta_m^t\}, w^{t,r})\|^2 && (31)\\
&= \mathbb{E}\|\frac{1}{R}\sum_{r=0}^{R-1}\left(\tilde{\nabla}_w\mathcal{R}(\{\theta_m^{t+1}\}, w^{t,r}) - \nabla_w\mathcal{R}(\{\theta_m^{t+1}\}, w^t)\right) + \nabla_w\mathcal{R}(\{\theta_m^{t+1}\}, w^t)\|^2 && (32)\\
&\leq 2\mathbb{E}\|\frac{1}{R}\sum_{r=0}^{R-1}\left(\tilde{\nabla}_w\mathcal{R}(\{\theta_m^{t+1}\}, w^{t,r}) - \nabla_w\mathcal{R}(\{\theta_m^{t+1}\}, w^t)\right)\|^2 + 2\mathbb{E}\|\nabla_w\mathcal{R}(\{\theta_m^{t+1}\}, w^t)\|^2 && \text{(from Lemma 13)}\\
&\leq 2\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\|\left(\tilde{\nabla}_w\mathcal{R}(\{\theta_m^{t+1}\}, w^{t,r}) - \nabla_w\mathcal{R}(\{\theta_m^{t+1}\}, w^t)\right)\|^2 + 2\mathbb{E}\|\nabla_w\mathcal{R}(\{\theta_m^{t+1}\}, w^t)\|^2 && \text{(Lemma 11)}\\
&\leq 4\sigma_R^2 + 32L_R^2\bar{\eta}_g^2(3G_R^2 + \frac{2\sigma_R^2}{R}) + 2G_R^2 = \psi_2. && (33)
\end{aligned}$$

Therefore,

$$\mathbb{E}[\mathcal{L}(w^{t+1}) - \mathcal{L}(\bar{\theta}^{t+1})] \leq \frac{\bar{\eta}_g}{2}G^2 + (\frac{\bar{\eta}_g}{2} + \frac{\bar{\eta}_g^2 L}{2})\psi_2. \tag{34}$$

□

Lemma 15 (From [34])**.**

$$\frac{1}{E}\mathbb{E}[\mathcal{L}(\bar{\theta}^{t+1}) - \mathcal{L}(w^t)] \leq \frac{1}{E}\sum_{e=0}^{E-1} -\frac{\eta_l}{2}\|\nabla\mathcal{L}(\bar{w}^{t,e})\|^2 + \frac{\eta_l^2 L\sigma^2}{2M} + 8\eta_l^3 E^2 L^2\bar{\gamma}^2. \tag{35}$$

Proof. We leverage the results from Equation (33) of [34] with $A_T = 0$ and $B_T = 1$[4], which are implied by Theorem 2 in [34]. □

[4] A_T and B_T are defined in [34].

Completing the proof of Theorem 3 Recall our main theorem

Theorem 3 (Convergence of PERADA global model)**.** *Let Assumptions 1 to 4 hold, and* $\eta_l = \frac{1}{EL\sqrt{T}}$, $\eta_g = \frac{1}{L_R RT}$, *denote* $\bar{w}^{t,e} = \frac{1}{M}\sum_{m=1}^{M}\theta_m^{t,e}$, *then the algorithm satisfies*

$$\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\frac{\mathbb{E}\|\nabla\mathcal{L}(\bar{w}^{t,e})\|^2}{ET} \leq \mathcal{O}\Big(\frac{L\Delta_{\mathcal{L}}+\psi_1}{\sqrt{T}} + \frac{\bar{\gamma}^2}{T} + \frac{L^2\psi_2}{T\sqrt{T}L_R^2 E}\Big),$$

where $\Delta_{\mathcal{L}} = \mathcal{L}(w^0) - \mathcal{L}(w^T)$, $\psi_1 = \frac{\sigma^2}{EM} + \frac{L(G^2+\psi_2)}{EL_R}$, *and* $\psi_2 = 4\sigma_R^2 + 32(3G_R^2 + \frac{2\sigma_R^2}{R})/T^2 + 2G_R^2$. *In particular,* $\bar{w}^{t+1,0} = w^t$ *and* $\bar{w}^{t+1,E-1} = \bar{\theta}^{t+1}$.

Proof. Combining Lemma 14 and Lemma 15,

$$\begin{aligned}\mathbb{E}[\mathcal{L}(w^{t+1}) - \mathcal{L}(w^t)] &= \mathbb{E}[\mathcal{L}(w^{t+1}) - \mathcal{L}(\bar{\theta}^t) + \mathcal{L}(\bar{\theta}^t) - \mathcal{L}(w^t)] &(36)\\ &\leq \frac{\bar{\eta}_g}{2}(G^2+\psi_2) + \frac{\bar{\eta}_g^2 L}{2}\psi_2 + \sum_{e=0}^{E-1} -\frac{\eta_l}{2}\|\nabla\mathcal{L}(\bar{w}^{t,e})\|^2 + \frac{E\eta_l^2 L\sigma^2}{2M} + 8\eta_l^3 E^3 L^2\bar{\gamma}^2. &(37)\end{aligned}$$

Rearrage the inequality and take $\frac{1}{T}\sum_{t=0}^{T-1}$ on both side. We get

$$\frac{1}{ET}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\|\nabla\mathcal{L}(\bar{w}^{t,e})\|^2 \leq \frac{2}{\eta_l ET}\big(\mathcal{L}(w^0) - \mathcal{L}(w^{T-1})\big) + \frac{\eta_l L\sigma^2}{M} + 16\eta_l^2 E^2 L^2\bar{\gamma}^2. + \frac{\bar{\eta}_g(G^2+\psi_2) + \bar{\eta}_g^2 L\psi_2}{E\eta_l} \tag{38}$$

Let $\eta_l = \frac{1}{LE\sqrt{T}}$ and $\eta_g = \frac{1}{L_R RT}$. Then,

$$\frac{1}{ET}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\|\nabla\mathcal{L}(\bar{w}^{t,e})\|^2 \leq \frac{2L}{\sqrt{T}}\big(\mathcal{L}(w^0) - \mathcal{L}(w^{T-1})\big) + \frac{\sigma^2}{EM\sqrt{T}} + 16\frac{\bar{\gamma}^2}{T} + \frac{L(G^2+\psi_2) + L^2\psi_2/L_R T}{EL_R\sqrt{T}} \tag{39}$$

□

D.3. Proofs for Personalized Model Convergence Guarantee in Theorem 4

Additional notations Let

$$\bar{\eta}_p = S\eta_p \tag{40}$$

Based on the update rules, we define $\delta_{v_m}^t$ as below, which capture the update of personalized model during client training.

$$v_m^{t+1} - v^t = -\bar{\eta}_p\delta_{v_m}^t \tag{41}$$

That is:

$$\delta_{v_m}^t := -\frac{1}{\eta_p S}(v_m^{t+1} - v^t) = \frac{1}{S}\sum_{s=0}^{S-1}\widetilde{\nabla}P_m(v_m^{t,s}, w^t) = \frac{1}{S}\sum_{s=0}^{S-1}\Big(\widetilde{\nabla}\mathcal{L}_m(v_m^{t,s}) + \lambda\left(v_m^{t,s} - w^t\right)\Big) \tag{42}$$

Proof Outline The goal is to bound the gradients of personalized models w.r.t the (Personal Obj), which is used to show that the trained models can converge to the stationary points:

$$\frac{1}{TS}\sum_{t=0}^{T-1}\sum_{s=0}^{S-1}\mathbb{E}\|\nabla_v P_m(v_m^{t,s}, w^t)\|^2 \tag{43}$$

Supporting lemmas We first introduce some supporting lemmas:

Lemma 16. *When* $\eta_p \le \frac{1}{L+\lambda}$*,*

$$\frac{1}{TS}\sum_{t=0}^{T-1}\sum_{s=0}^{S-1}\|\nabla_v P_m(v_m^{t,s}, w^t)\|^2 \le \frac{2(P_m(v^0, w^0) - P_m(v^T, w^T))}{\eta_p TS} + (L+\lambda)\eta_p\delta^2 + \frac{1}{TS}\sum_{t=0}^{T-1}\frac{G_P\mathbb{E}\|w^{t+1} - w^t\|_2}{\eta_p} \tag{44}$$

Proof. Let $\eta_p \le \frac{1}{L+\lambda}$.

$$\begin{aligned}
\mathbb{E}[P_m(v_m^{t,s+1}, w^t) - P_m(v_m^{t,s}, w^t)] &\le \mathbb{E}\langle \nabla_v P_m(v_m^{t,s}, w^t), -\eta_p\tilde{\nabla}_v P_m(v_m^{t,s}, w^t)\rangle + \frac{(L+\lambda)\eta_p^2}{2}\mathbb{E}\|\tilde{\nabla}_v P_m(v_m^{(t,s)}, w^t)\|^2 \\
&\qquad\qquad \text{(Assumption 1)} \\
&\le -\eta_p\|\nabla_v P_m(v_m^{t,s}, w^t)\|^2 + \frac{(L+\lambda)\eta_p^2}{2}(\sigma^2 + \|\nabla_v P_m(v_m^{t,s} - w^t)\|^2) \qquad (45) \\
&\le -\frac{\eta_p}{2}\|\nabla_v P_m(v_m^{t,s}, w^t)\|^2 + \frac{(L+\lambda)\eta_p^2\sigma^2}{2}. \qquad \text{(By } \eta_p \le \tfrac{1}{L+\lambda}\text{)}
\end{aligned}$$

Taking a summation $\sum_{s=0}^{S-1}$,

$$\mathbb{E}[P(v_m^{t+1}, w^t) - P(v_m^t, w^t)] \le -\frac{\eta_p}{2}\sum_{s=0}^{S-1}\|\nabla_v P_m(v_m^{t,s}, w^t)\|^2 + \frac{(L+\lambda)S\eta_p^2\sigma^2}{2}. \tag{46}$$

We next bound $P_m(v_m^{t+1}, w^{t+1}) - P_m(v_m^{t+1}, w^t)$. Since bounded gradient implies Lipschitz function, P_m is G_P Lipschitz in terms of w, i.e.,

$$\mathbb{E}[P_m(v_m^{t+1}, w^{t+1}) - P_m(v_m^{t+1}, w^t)] \le G_p\mathbb{E}\|w^{t+1} - w^t\|_2. \qquad \text{(Assumption 4)}$$

Combine the two statements, rearrange, and take the sum $\sum_{t=0}^{T-1}$ on both side:

$$\frac{1}{TS}\sum_{t=0}^{T-1}\sum_{s=0}^{S-1}\mathbb{E}\|\nabla_v P_m(v_m^{t,s}, w^t)\|^2 \le \frac{2(P_m(v^0, w^0) - P_m(v^T, w^T))}{\eta_p TS} + (L+\lambda)\eta_p\delta^2 + \frac{1}{TS}\sum_{t=0}^{T-1}\frac{2G_P\mathbb{E}\|w^{t+1} - w^t\|_2}{\eta_p} \tag{47}$$

□

Lemma 17. *When* $\eta_l = \frac{1}{EL\sqrt{T}}$*,*

$$\mathbb{E}\|w^{t+1} - w^t\|^2 \le 8\eta_l^2\sigma^2 + \frac{\eta_l^2\phi_1}{T} + 4\eta_l^2\mathbb{E}\|\nabla\mathcal{L}(w^t)\|^2 + 2\eta_g^2R^2G_R^2. \tag{48}$$

where $\phi_1 = 64(3\bar{\gamma} + \frac{2\sigma^2}{E})$*.*

Proof. By definition

$$\begin{aligned}
\mathbb{E}\|w^{t+1} - w^t\|^2 &= \mathbb{E}\|\eta_l E\frac{1}{M}\sum_{m=1}^{M} g_m^t + \eta_g R g^t\|^2 \\
&\le 2\eta_l^2E^2\mathbb{E}\|\frac{1}{M}\sum_{m=1}^{M} g_m^t\|^2 + 2\eta_g^2R^2\mathbb{E}\|g^t\|^2 \qquad (49)
\end{aligned}$$

For the first term,

$$
\begin{aligned}
\mathbb{E}\|\frac{1}{M}\sum_m g_m^t\|^2 &= \mathbb{E}\|\frac{1}{M}\sum_m \Big(\frac{1}{E}\sum_{e=0}^{E-1}\tilde{\nabla}\mathcal{L}_m(\theta_m^{t,e}) - \nabla\mathcal{L}_m(w^t)\Big) + \nabla\mathcal{L}(w^t)\|^2 \\
&\leq \frac{2}{EM}\sum_{m=1}^{M}\sum_{e=0}^{E-1}\mathbb{E}\|\tilde{\nabla}\mathcal{L}_m(\theta_m^{t,e}) - \nabla\mathcal{L}_m(w^t)\|^2 + 2\mathbb{E}\|\nabla\mathcal{L}(w^t)\|^2 \\
&\leq 4\sigma^2 + 64\eta_l^2E^2L^2(3\bar{\gamma} + 3\mathbb{E}\left\|\nabla\mathcal{L}(w^t)\right\|^2 + \frac{2\sigma^2}{E}) + 2\mathbb{E}\|\nabla\mathcal{L}(w^t)\|^2 && \text{(Lemma 12)} \\
&= 4\sigma^2 + 64\frac{1}{T}(3\bar{\gamma} + 3\mathbb{E}\left\|\nabla\mathcal{L}(w^t)\right\|^2 + \frac{2\sigma^2}{E}) + 2\mathbb{E}\|\nabla\mathcal{L}(w^t)\|^2 \\
&= 4\sigma^2 + \frac{1}{T}\phi_1 + (2 + \frac{192}{T})\mathbb{E}\|\nabla\mathcal{L}(w^t)\|^2
\end{aligned}
$$

For the second term, recall Equation (33), then we have

$$
\mathbb{E}\|g^t\|^2 \leq \psi_2 \tag{50}
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}\|w^{t+1} - w^t\|^2 &\leq 2\eta_l^2E^2\mathbb{E}\|\frac{1}{M}\sum_{m=1}^{M} g_m^t\|^2 + 2\eta_g^2R^2(\sigma_R^2 + G_R^2) \\
&= 8\eta_l^2\sigma^2 + \frac{\eta_l^2\phi_1}{T} + 4\eta_l^2(1 + \frac{96}{T})\mathbb{E}\|\nabla\mathcal{L}(w^t)\|^2 + 2\eta_g^2R^2\psi_2.
\end{aligned}
$$

□

Completing the proof of Theorem 4 Recall Theorem 4:

Theorem 4 (Convergence of PERADA personalized model)**.** *When* $\eta_p = \frac{1}{(L+\lambda)\sqrt{T}}$, $\eta_l = \frac{1}{EL\sqrt{T}}$, $\eta_g = \frac{1}{L_RRT}$, *then the algorithm satisfies:*

$$
\frac{1}{TS}\sum_{t=0}^{T-1}\sum_{s=0}^{S-1}\mathbb{E}\|\nabla_vP_m(v_m^{t,s}, w^t)\|^2 \leq \mathcal{O}\Big(\frac{(L+\lambda)\Delta_{P_m} + \phi_2}{\sqrt{T}S} + \frac{G_P(L+\lambda)(L\Delta_{\mathcal{L}} + \psi_1)^{1/2}}{T^{1/4}L\sqrt{E}S} + \frac{G_P(L+\lambda)\sqrt{\psi_2}}{T^{3/4}L_RES} + \frac{G_P(L+\lambda)\sigma}{LES}\Big)
$$

where $\Delta_{P_m} = P_m(v_m^0, w_m^0) - P_m(v_m^t, w^t)$, $\phi_1 = 64(3\bar{\gamma} + \frac{2\sigma^2}{E})$, $\phi_2 = S\sigma^2 + \frac{\sqrt{\phi_1}G_P(L+\lambda)}{LE} + \sqrt{\psi_2}G_P(L+\lambda) + \frac{G_P\bar{\gamma}(L+\lambda)}{L\sqrt{E}}$. ψ_1, ψ_2 *are defined the same as in Theorem 3.*

Next, we combine Lemma 16 and Lemma 17 to prove the above theorem.

Proof. From Lemma 16,

$$
\frac{1}{TS}\sum_{t=0}^{T-1}\sum_{s=0}^{S-1}\mathbb{E}\|\nabla_vP_m(v_m^{t,s}, w^t)\|^2 \leq \frac{2\Delta_{P_m}}{\eta_pTS} + (L+\lambda)\eta_p\sigma^2 + \frac{L+\lambda}{\sqrt{T}S}\sum_{t=0}^{T-1}2G_P\mathbb{E}\|w^{t+1} - w^t\|_2 \tag{51}
$$

Expand the last term according to Lemma 17.

$$
\begin{aligned}
\frac{1}{\sqrt{T}}\sum_{t=0}^{T-1}\mathbb{E}\|w^{t+1}-w^t\| &\leq \mathbb{E}\sqrt{\sum_{t=0}^{T-1}\|w^{t+1}-w^t\|^2} && \text{(Taking square root for Lemma 11)}\\
&\leq \sqrt{\sum_{t=0}^{T-1}\mathbb{E}\|w^{t+1}-w^t\|^2}\\
& && \text{(Jensen's inequality } \mathbb{E}[f(x)] \leq f(\mathbb{E}[x]) \text{ for the concave function } f(x))\\
&= \sqrt{8\sigma^2\eta_l^2T+\eta_l^2\phi_1+4\sum_{t=0}^{T-1}\eta_l^2\mathbb{E}\|\nabla\mathcal{L}(w^t)\|^2+2\eta_g^2R^2\psi_2T}\\
&\leq \mathcal{O}(\sigma\sqrt{T}\eta_l)+\eta_l\sqrt{\phi_1}+2\eta_l\cdot\sqrt{\sum_{t=0}^{T}\sum_{e=0}^{E-1}\mathbb{E}\|\nabla\mathcal{L}(\bar{w}^{t,e})\|^2}+\mathcal{O}(\eta_gR\sqrt{\psi_2}\sqrt{T})\\
& && (\sqrt{\textstyle\sum_{i=1}^n x_i^2} \leq \textstyle\sum_{i=1}^n x_i \text{ for } x_i \geq 0)\\
&\leq \mathcal{O}(\frac{\sigma}{LE})+\frac{\sqrt{\phi_1}}{LE\sqrt{T}}+\frac{2}{L\sqrt{E}}\mathcal{O}\Big(T^{-1/4}(L\Delta_{\mathcal{L}}+\psi_1)^{1/2}+\frac{\bar{\gamma}}{\sqrt{T}}+T^{-3/4}\frac{L\sqrt{\psi_2}}{L_R\sqrt{E}}\Big)+\frac{\sqrt{\psi_2}}{\sqrt{T}}.
\end{aligned}
$$

$$
\begin{aligned}
&\frac{1}{TS}\sum_{t=0}^{T-1}\sum_{s=0}^{S-1}\mathbb{E}\|\nabla_vP_m(v_m^{t,s},w^t)\|^2\\
&\leq \frac{2\Delta_{P_m}(L+\lambda)}{\sqrt{T}S}+\frac{\sigma^2}{\sqrt{T}}\\
&+\frac{(L+\lambda)2G_P}{S}\left(\mathcal{O}(\frac{\sigma}{LE})+\mathcal{O}\left((\frac{\sqrt{\phi_1}}{LE}+\sqrt{\psi_2}+\frac{2\bar{\gamma}}{L\sqrt{E}})\frac{1}{\sqrt{T}}\right)+\frac{2}{L\sqrt{E}}\mathcal{O}\Big(T^{-1/4}(L\Delta_{\mathcal{L}}+\psi_1)^{1/2}+T^{-3/4}\frac{L\sqrt{\psi_2}}{L_R\sqrt{E}}\Big)\right)\\
&=\mathcal{O}(\frac{(L+\lambda)\Delta_{P_m}+\phi_2}{\sqrt{T}S})+\frac{(L+\lambda)2G_P}{S}\left(\mathcal{O}(\frac{\sigma}{LE})+\frac{2}{L\sqrt{E}}\mathcal{O}\Big(T^{-1/4}(L\Delta_{\mathcal{L}}+\psi_1)^{1/2}+T^{-3/4}\frac{L\sqrt{\psi_2}}{L_R\sqrt{E}}\Big)\right)
\end{aligned}
$$

□